The Thesis Committee for Reem A Goudy
certifies that this is the approved version of the following Thesis:

Unsupervised Fine-Tuning Data Selection for ASR Using Self-Supervised Speech Models

APPROVED BY

SUPERVISING COMMITTEE:

David Harwath, Supervisor

Eunsol Choi

Unsupervised Fine-Tuning Data Selection for ASR Using Self-Supervised Speech Models

by

Reem A Goudy

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computer Science

The University of Texas at Austin

December 2022

Dedication

Dedicated to my supportive parents.

Acknowledgments

I wish to thank the multitudes of people who helped me throughout my life and supported me in my learning journey. First, I'd like to thank my parents who believe in me and are always encouraging and supportive. Second, I'd like to thank all the professors at Cairo University who have spent considerable effort in designing their courses and in growing our passion for the field. I'd like to give special thanks to my graduation project adviser Dr. Ahmed S. Kaseb who has been always supportive of our ideas and who has spent considerable effort in teaching us how to improve our writing in academic papers. Moreover, I'd like to thank my manager at work Dr. Ahmed Tawfik, and my mentor Alaa ElShafaey for the tremendous support that they offered during my master's journey. Finally, I'd like to thank professor David Harwath for his supervision and guidance. I really enjoyed our 1:1 meetings and how we brainstormed multiple ideas during these meetings.

Abstract

Unsupervised Fine-Tuning Data Selection for ASR Using Self-Supervised Speech Models

Reem A Goudy, M.S.C.S
The University of Texas at Austin, 2022

Supervisor: David Harwath

Self-supervised learning (SSL) has been able to leverage unlabeled data to boost the performance of automatic speech recognition (ASR) models when we have access to only a small amount of transcribed speech data. However, this raises the question of which subset of the available unlabeled data should be selected for transcription. Our work investigates different unsupervised data selection techniques for fine-tuning the HuBERT model under a limited transcription budget. We investigate the impact of speaker diversity, gender bias, and topic diversity on the downstream ASR performance. We also devise two novel techniques for unsupervised data selection: pre-training loss based data selection and the perplexity of byte pair encoded clustered units (PBPE) and we show how these techniques compare to pure random data selection.

Finally, we analyze the correlations between the inherent characteristics of the selected fine-tuning subsets as well as how these characteristics correlate with the resultant word error rate. We demonstrate the importance of token diversity, speaker diversity, and topic diversity in achieving the best performance in terms of WER.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Self-supervised speech recognition models like wav2vec 2.0 and HuBERT [2, 19] have achieved very low WER when pre-trained on a large dataset of untranscribed speech and fine-tuned on as little as 1 hour of transcribed data. This motivates using these models for automatic speech recognition for low-resource scenarios where we may have access to a moderate to large amount of untranscribed speech but a finite budget is available for data transcription. However, this raises the question of how to optimally choose which subset of the data should be transcribed for fine-tuning the model. Using small amounts of data in fine-tuning is associated with the risk of having high variance in the WER at test time, depending on the characteristics of the subset selected for fine-tuning. Moreover, the fact that the data selection pool is untranscribed implies the necessity of devising unsupervised techniques for data selection that do not rely on the existence of transcriptions. Our goal in this work is to investigate different selection criteria for choosing the fine-tuning subset, and their effect on downstream ASR performance. In our setup, we assume that we have a large pool of unlabeled in-domain data and a limited transcription budget, e.g. 10 hours. We need to select a subset of this data pool to transcribe and use for fine-tuning the model. We probe the impact of speaker diversity, gender bias and topic diversity on the model performance.

Moreover, we devise two novel techniques for unsupervised data selection: Pre-trained loss based data selection and Perplexity of byte pair encoded clustered units (PBPE) and we show how these techniques compare to pure random data selection.

The main contributions of this thesis can be summarized as follows:

- Studying the impact of different selection criteria for data used in fine-tuning pre-trained speech models like HuBERT on the downstream automatic speech recognition task when a limited budget is available for transcription.
- Introducing two new techniques for unsupervised fine-tuning data selection for the HuBERT model. The first technique is based on the pre-training loss of HuBERT, and the second one is based on the perplexity computed over BPE HuBERT clustered units.
- Conducting deep analysis on how the different properties of the selected fine-tuning subsets correlate with the final WER obtained on the test sets.
- To the best of our knowledge, this is the first work that studies and analyzes the impact of the different variables associated with data selection for fine-tuning (e.g.,number of speakers, number of vocabulary words,total number of utterances, etc.) on the downstream performance of the HuBERT model when selecting a limited subset of data for fine-tuning from a pool of unlabeled in-domain data.

The thesis is organized as follows. In chapter 2, we present an overview of the speech recognition pipeline and highlight some related work that has been conducted for data selection in both the semi-supervised as well as the self-supervised learning paradigms. In chapter 3, we present the different fine-tuning data selection criteria that we investigate in this work. In addition to this, we describe the novel data selection criteria that we devise. In chapter 4, we describe our experimental setup, present the results of our work, and provide a deep analysis of the impact of the different variables associated with the fine-tuning data selection on the downstream word error rate on the test sets that we use in this work. Finally, we highlight our conclusions in chapter 5 and propose directions for future work.

Chapter 2: Background and Related Work

In this chapter, we provide a brief overview of the history and techniques used in automatic speech recognition (ASR). The ASR pipeline has evolved from using Hidden Markov Model (HMM) trained in a supervised manner to employing large neural network architectures for building speech recognition models in either a hybrid or end-to-end manner. Various methods have been investigated in training neural networks for the ASR task. Supervised learning methods have been explored in that regard, and then the need for large amounts of data and the desire to build larger models with more learnable parameters have driven the exploration of other training techniques as semi-supervised learning and self-supervised learning. Using these techniques, the training pipeline can benefit from the abundance of large amounts of unlabeled data and leverage that for improving the ASR models. However, the sensitivity of the neural networks to the data that they are fed during training and the deployment of the ASR models in systems operating in specific domains among other factors have driven research in data selection and filtration to achieve better performance. In section 2.1, we give an overview of the automatic speech recognition pipeline. Furthermore, section 2.2 highlights the techniques used in training the ASR models and some work related to data selection techniques in both the semi-supervised and the self-supervised

learning paradigms. The work presented in this thesis aims to investigate unsupervised data selection methods for fine-tuning speech models trained in a self-supervised manner for the ASR task.

2.1 Overview on Automatic Speech Recognition Pipeline

The automatic Speech Recognition (ASR) pipeline has evolved tremendously in the last decade. The typical ASR pipeline comprises front-end processing which is mainly about feature extraction, acoustic modeling, language modeling, pronunciation modeling and decoding. Scientists have studied different representations for speech signals throughout the years, and have deeply investigated the use of different features that capture the information in the speech signal. The most popular features that are being used in ASR to this time are: mel frequency cepstral coefficients (MFCC), Log-Mel filter banks (LFB), discrete wavelet transforms (DWT), linear predictive coding (LPC), and perceptual linear predictive coefficients (PLP) [32, 42]. In addition to feature engineering and extraction, massive work has been done on the modeling side. As the speech signal is a temporal complex signal that exhibits a lot of variability based on the content, the speaker and the background noise, building models that are robust to such conditions, and are simultaneously able to capture the sequential information in a given context has been crucial. More formally, the main task in speech recognition is to find the most probable sequence of words that account for a specific utterance, i.e. that maximize this probability $P(W|X)$, where W is the sequence of words, and X is the sequence

of features encoding a particular utterance. This probability can be re-written using Bayes rule as:

$$P(WjX) = \frac{P(XjW)P(W)}{P(X)} \;/\; P(XjW)P(W) \tag{2.1}$$

and the problem is reduced to finding the optimal sequence of words W^* such that

$$W^* = argmax_W P(XjW)P(W) \tag{2.2}$$

where $P(XjW)$ is the likelihood given by the acoustic model and represents how likely it is for a sequence of words to correspond to a given utterance, and $P(W)$ is given by the language model and represents how likely it is for a sequence of words to be observed in the language in general.

2.1.1 Acoustic Model

Acoustic modeling is tackled as a multi-class classification problem, where it is required to find a class label for each feature vector, taking its context into consideration. The inherent sequential nature of a speech signal makes it hard to ignore the temporal dependencies and hence extremely implausible to find a class label for each feature vector in isolation. For each feature vector, presumably extracted from a stationary segment of the speech signal, the class label is the corresponding unit of sound that accounts for the observed feature vector. Using the phonemes as class labels often results in high variability within the same class, as the acoustic realization of each phoneme also depends on the context in which it occurs. As a result, context-dependent phones (also known as triphones) have also been studied as more

fine-grained class labels since they assign different labels to the same phone depending on its acoustic context. Accordingly, each word is represented as a sequence of phonemes/triphones and a pronunciation model or lexicon is used to provide that mapping.

Due to its ability to model temporal dependencies resulting in a sequence of observations, the HMM has been one of the most popular statistical models used in ASR for years [12]. For each class (phoneme or triphone), an HMM is trained using feature vectors corresponding to that class which are extracted from multiple speech signals. The typical HMM used in this context consists of three states, in addition to a start state and an end state to account for the acoustic variability and duration variance observed within the same class label. Gaussian Mixture models (GMMs) have been used to model the density functions associated with the HMM states, due to their ability to model complex distributions as a weighted sum of Gaussian distributions, which in turn makes them suitable for the multi-modal distributions of the phonetic features. As this can result in a large number of parameters associated with the Gaussian components, state tying has been used to achieve parameter sharing between similar states. Despite this approach being a successful solution for the acoustic modeling problem, the HMM has unrealistic conditional independence assumptions, not applicable to the nature of the speech signals, but put in place to make the training and decoding tractable. For this reason, scientists have looked into neural networks as an alternative to HMMs. In 2009, Abdel-rahman Mohamed et al [31] proposed using Deep Belief Networks

(DBNs) for acoustic modeling and were able to achieve 23% phone error rate (PER) on the Timit test set beating the bayesian triphone HMM [30] which had a PER of 25.6% on the same test set. However, instead of giving up on the HMM approach and replacing it with neural networks, the use of neural networks has been explored as an alternative to GMMs [43] in the HMM framework, and in 2011, Frank Seide et al [46] introduced Context-Dependent Deep-Neural-Network HMMs (CD-DNN-HMMs) that achieved an impressive 33% relative improvement on Switchboard test set compared to GMM-HMM. This result highly encouraged the use of neural networks with HMMs in acoustic modeling and triggered explorations of different neural architectures within the same framework. This has also been accompanied by drastic improvements in hardware and an increase in the amounts of available training data, making it more feasible to train neural networks for acoustic modeling. Among the other architectures that have been explored in the HMM framework are: deep belief networks (DBN-DNN)[18], recurrent neural networks (RNNs) and their variants as long short term memory (LSTM) and gated recurrent units (GRU) [23], bi-directional long short term memory networks (BiLSTMs) [15, 55], time-delayed Neural networks (TDNNs) [39], convolutional neural networks (CNNs) [38, 44] , and transformers [50].

2.1.2 Language Model

The language model assigns scores to each sentence depending on how likely it is to occur in the language. In ASR, n-gram language models are

commonly used, as they can be easily defined as weighted finite-transducers (WFST) and hence composed with the rest of the components of the ASR system which are the HMM finite state machine (FST) and the lexicon FST to form the decoding graph [6]. One other common use for language models in ASR is rescoring the n-best hypotheses that result from decoding. In this context, the rescoring language model could be an n-gram language model or a neural language model [13].

2.1.3 End to End Architectures in ASR

In parallel to the research conducted on hybrid acoustic models which is described in section 2.1.1, scientists have investigated the employment of a single neural model for ASR; a model that is fed with raw speech or featurized speech and is able to produce the text transcription in an end to end manner. In end-to-end speech recognition, the alignment between the input frames and the phonetic transcription is not required, and neither is the pronunciation lexicon. This represents a major simplification in the ASR pipeline, which makes it an attractive direction for the next generation of speech models. However, this comes at a cost of an increase in the computation requirements, as well as the required data to achieve good performance using this technique. In 2014, Hannun et Al [17] proved the technique to be very promising by achieving 16% WER on the full SwitchBoard test set, which was a competitive result to existing hybrid systems at that time [26, 44, 45]. In that work, they trained an RNN model using CTC loss and managed to scale the labeled training data

to 5000 hours of read speech from 9600 speakers. Since then, the end-to-end approach in ASR has witnessed several explorations in terms of the network architecture, the training loss function, and the training techniques. In addition to the common supervised training pipeline, semi-supervised training, as well as self-supervised learning, have been investigated in ASR in order to leverage larger amounts of training data to train even larger neural architectures.

2.2 Overview on the Main Training Techniques for an ASR Model

In the previous section, we shed light on hybrid and end-to-end models in ASR. For a long time, supervised learning has been the main method for training these models. Hence, the abundance of labeled data has been a limiting factor in achieving good performance on downstream ASR, especially on low-resource languages. This has led to the implementation of other techniques such as semi-supervised learning and self-supervised learning that aim to leverage unlabeled data which is naturally available in much larger quantities.

2.2.1 Supervised Learning in ASR

Supervised learning has been the standard method of training speech recognition models for a long time. It requires access to large amounts of transcribed speech, and in the context of hybrid models mandates the alignment of speech to its corresponding transcription. In spite of this approach

achieving good performance in ASR in terms of WER, it isn't easily scalable to large network architectures, which require large amounts of data to avoid the over-fitting phenomenon. The need for more labeled data incurs an extra cost of transcription and exposes limitations to scaling across multiple languages, especially the low resource ones. In addition to this, supervised training exhibits more robustness and generalization when trained on large amounts of data from different sources as shown by the authors in [7, 28]. In [7], the authors mix a total of 5,140 hours of labeled English data in their model training. This amount of data can't be easily collected for all the languages and it's small when compared to the amount of unlabeled data that can be obtained. One solution to scale the data in supervised learning is to relax the requirement of golden transcription that undergoes human validation and leverage automated pipelines to obtain more labeled data. This sacrifices quality for quantity and introduces a learning technique known as weakly supervised learning. [41] introduces whisper, which is a demonstration of weakly supervised learning at scale. In this work, 680,000 labeled hours are used in training a sequence to sequence model that operates on multiple speech related tasks. This data is scraped from the internet and is the result of applying a set of heuristic filtering steps in order to pick the cleanest data to use for training. The aim of this work is to show how robust supervised pre-training can be in a zero-shot setup. While achieving robustness and high generalization on multiple domains has been a long sought goal in ASR, specializing the ASR model on a particular target domain has been another active research area with more promise in

terms of feasibility and good performance. In that direction, data selection techniques have been studied to improve the performance of supervised models through domain adaptation. This entails leveraging labeled out-of-domain data by selecting utterances that are similar to a particular target domain.

2.2.2 Semi-supervised Learning in ASR

Semi-supervised learning has been considered to increase the amount of data that can be used to train neural networks for speech recognition. It is a means by which we can leverage unlabeled data to train a model in a supervised way. One major direction in semi-supervised learning is self-training. In self-training, a teacher model is trained in a supervised manner using labeled data, and is used to generate predictions for unlabeled data. The probability distributions over the predicted class labels are used as soft labels, or the maximum scoring class labels are used as hard labels for the originally unlabeled samples. Since these labels are artificially generated, they are known as pseudo-labels. A student model is then trained on both the labeled and pseudo-labeled examples in a supervised manner. One flavour of self-training is called noisy student training (NST) and involves injecting noise through drop-out or data augmentation during training the student. This technique has been initially studied in an image classification problem, and significantly improved the performance on ImageNet test set [53]. It has then been applied in speech recognition, where SpecAugment [36] has been used as a means of generating noise during the student training [37, 51, 56]. Iterative pseudo-labeling has also

been considered as a means of obtaining more refined pseudo-labels [40, 54]. In this technique, the pseudo-labels are iteratively refined as the performance of the model used in the decoding is improved across the iterations. In [54], the authors only label one subset of the unlabeled data in each iteration, and fine-tune the existing model on this subset rather than starting from scratch. Semi-supervised data has also been studied when the available unlabeled data is out-of-domain. In [4], the author uses out-of-domain data to improve low resource ASR performance. A seed model is trained on the available domain data and then used to transcribe the out-of-domain data. In this work, the WER improvements obtained when training a model with the pooled in-domain and out-of-domain data, then fine-tuning it with the in-domain data, are higher than those obtained by just training on the pooled data.

2.2.2.1 Data Selection Techniques in Semi-supervised Learning

Since semi-supervised learning involves artificially generating data by using a seed model to transcribe unlabeled data, it is certainly entitled to suffer from transcription errors, and hence the pseudo-labeled data quality becomes a major concern. Accordingly, data selection and filtration have been extensively studied in the semi-supervised training paradigm, with the most common technique being data filtration based on confidence scores [1, 9, 21, 48]. Kahn et al [21] revisit self-training in the context of end to end (E2E) models. They use an ensemble of four models to ensure pseudo-label diversity and investigate the use of heuristic based mechanisms relevant to sequence to

sequence models for data filtration. As looping and early generation of end of sentence (EOS) tokens are two common phenomena occuring in the context of sequence to sequence models, they filter the utterances that either have a repetition of a particular n-gram more than a specific number of times, as well as utterances in which the EOS token has been generated with a low probability. In addition to this, they combine their heuristic based filtration techniques with confidence based measures. In [48], the authors use utterance level confidence scores for data selection. They hypothesize that different domains are likely to exhibit different performance, and hence different thresholds for filtration. Accordingly, they used a natural language understanding (NLU) system to find the domain of each utterance. They investigate fine-tuning on a single domain vs fine-tuning on the combined domains. They show that sampling from a particular domain improves the performance on that domain without significant degradation on the other domains. Wotherspoon et al [52] show that they can achieve good WER results when they have transcribed out-of-domain data and untranscribed in-domain data if they carefully select the in-domain data for which they generate the pseudo-labels. They train a model with labeled out-of-domain data in a supervised manner, then they use it to transcribe the in-domain data. They use phone confidence based data selection criterion and achieve the best results on the target domain when selecting 3% of their unlabeled data subset. In their experiments, larger selections degrade the performance. They hypothesize that it is better to use a smaller subset of the pseudo-labeled in-domain data because the transcripts

originally generated by a model trained on out-of-domain data are expected to have more errors in case of a large domain shift.

The most obvious drawback of data selection based on confidence scores is its sensitivity to the choice of the confidence threshold, leading to a trade-off between data quality and quantity. For this reason, other tracks for data selection in semi-supervised training have also been investigated. In [11], the authors used an ensemble technique that relies on the agreement of models trained on different dialects of the same language in selecting high quality utterances. They hypothesize that models trained on different dialects of the same language that are close to each other can result in diversity in the output mistakes. Consequently, if these models agree on a transcription of a particular utterance, then it is likely to be correct and is thus selected. However, this approach has the drawback of discarding utterances with dialect-specific keywords.

2.2.3 Self-supervised Learning in ASR

Self-supervised learning is an alternative method to self-training that is also designed to leverage large amounts of unlabeled data to improve the downstream performance of ASR models. Instead of using a seed model to generate pseudo-labels for the unlabeled data as in the semi-supervised training paradigm, this approach relies on using various pre-training objectives that don't rely on the existence of transcriptions for the available data, but rather utilize the internal structure of the data. In this paradigm, the train-

ing procedure is divided into two stages: pre-training and fine-tuning. In the pre-training stage, all the unlabeled data is used to train the model in an unsupervised way. After, that the model is fine-tuned on a downstream task (like ASR) using a much smaller amount of the available transcribed data. The most popular models that utilize this technique are wav2vec 2.0 [2], HuBERT [19] and wavLM [10].

2.2.3.1 Wav2vec 2.0 Model

Wav2vec 2.0 is the first model to show that learning speech representations from raw audio and fine-tuning on transcribed data can outperform semi-supervised training methods. The model is composed of a feature encoder that gets raw speech as input and outputs latent speech representations. The latent speech representations are contextualized by passing them through a transformer model. In the same time, the latent speech representations pass through a quantization module, which uses product quantization to choose quantized representations from multiple codebooks. During pre-training, a percentage of the latent representations that are output from the feature encoder are masked, then the model needs to identify the quantized representation corresponding to each masked step among a set of k distractors that are uniformly sampled from other quantized representations in the same utterance using a contrastive loss objective. For fine-tuning the pre-trained model for ASR, a randomly initialized linear layer is added on top of the transformer to project the output into a number of classes which represent the target units

(letters or subwords) and a connectionist temporal classification (CTC) loss is used for fine-tuning the model on transcribed speech. It is demonstrated that wav2vec 2.0 can be fine-tuned on as little as 1 hour of transcribed data and produce competitive performance [2], and hence making self-supervised learning an intriguing technique for low resource tasks that lack the appropriate amount of labeled data required to obtain decent performance in the supervised paradigm.

2.2.3.2 HuBERT Model

The HuBERT model follows the same architecture as wav2vec 2.0. It has a convolutional waveform encoder, which downsamples the audio by 320x and generates the feature sequence at 20 ms frame rate for audio sampled at 16 KHZ. Spans of length l of encoded features are then masked, where the start of the span is selected for masking with probability $p\%$. The masked sequence is then passed through a BERT like transformer encoder which generates the contextualized embeddings. The contextualized embeddings are projected and their cosine similarity with the codeword embeddings are computed. Finally, a softmax function is applied to find the output distribution over the codewords. To pre-train HuBERT, an offline K-means clustering step is used to generate the training labels. First, the MFCC features are clustered to generate the labels and the model is pre-trained using cross entropy loss over the masked frames for a specified number of training steps. The cluster labels are then refined, by applying the clustering step over the learned latent representations

that are extracted from an intermediate layer in the transformer stack to produce better labels and then the training is repeated for more iterations. Similar to the result demonstrated in wav2vec 2.0, HuBERT shows impressively low WER results when fine-tuned with 1 hour or 10 hours of transcribed speech [19]. The experiments in this thesis are based on the HuBERT model.

2.2.3.3 Data Selection in Self-supervised Learning

In line with the impressive results obtained by using very small amounts of transcribed data in the self-supervised learning paradigm, it becomes very important to analyze the impact of the choice of the pre-training data and the fine-tuning data on the downstream performance in both the in-domain and the out-of-domain setups. The huge amounts of data that are utilized in pre-training the models imply the large dependence on computing power, which is unlikely to be available for the research community to build similar models. This gives rise to multiple questions:

- **What is the impact of the domain shift between pre-training and fine-tuning?**

  If it is difficult to replicate the pre-trained models for each domain, either due to the lack of computing resources or the shortage of unlabeled data in that particular domain, it is important to know the impact of fine-tuning an existing pre-trained model on transcribed data from a different target domain. It can also be the case that we have a myriad of unlabeled data from a particular target domain, and enough compute

resources for pre-training the model on that data, but we only have out-of-domain transcribed data, which also encourages investigating the impact of domain shift between pre-training and fine-tuning. The works in [20, 24] study this problem. In [20], the authors show that better results are obtained when the pre-training data includes in-domain data.

- **Is a very large amount of unlabeled data required in the pre-training stage?**

  If we are going to fine-tune the pre-trained model on a target domain, then it might be the case that we can select a certain portion of the unlabeled data to use for pre-training in such a way that it is similar to the intended target domain without incurring performance loss after fine-tuning. This has the advantage of decreasing the amount of compute resources required for pre-training by using a smaller subset without suffering degradation in performance. [29] uses a contrastive data selection method applied to the learned discrete tokens for selecting data for pre-training an ASR model. They show a significant improvement when selecting pre-training data that is matched to the target domain compared to pre-training using a full data set of 1 million youtube hours.

- **Is there a way to select the fine-tuning data that guarantees optimal performance?**

  In situations in which the transcription budget is limited (for example, 1 hour or 10 hours), it is highly plausible that the performance on the

downstream speech recognition task exhibits high variance based on how the fine-tuning data is selected. This implies the importance of devising unsupervised techniques for fine-tuning data selection that guarantee optimal performance in test time. The work in this thesis falls in this category, where we study the impact of the different data selection criteria on the downstream WER of an ASR model when we have a limited transcription budget. The experiments we conduct assume access to a large pool of unlabeled data from a particular target domain, and the task is to select only 1 hour or 10 hours of data to transcribe and use for fine-tuning a model that is pre-trained on the same domain. We show the impact of speaker diversity, content diversity, and gender bias on the WER. Furthermore, we devise two novel techniques for unsupervised data selection: perplexity of byte pair encoded clustered units and pre-training loss based data selection. To the best of our knowledge, this is the first work that studies the different selection techniques for fine-tuning data subsets in the low resource setup for the HuBERT model.

In this chapter, we gave a brief overview of the automatic speech recognition pipeline and the techniques that are used for training ASR models. In addition to this, we highlighted some of the data selection techniques used in semi-supervised and self-supervised learning. We showed that our work lies under unsupervised techniques for fine-tuning data selection for self-supervised

speech recognition models. In the next chapter, we discuss our proposed criteria for fine-tuning data selection in more detail.

Chapter 3: Proposed Selection Criteria

In this chapter, we describe the different selection criteria that we apply for selecting the fine-tuning data for the HuBERT base model. We try several criteria for data selection using the Librispeech [34] dataset with a limited transcription budget of 1 hour or 10 hours of speech and compare these criteria to pure random selection. We investigate the effect of speaker diversity by forcing a specified number of speakers into the fine-tuning subset. Moreover, we probe the impact of gender bias by selecting subsets with either female or male speakers only. We examine the impact of topic diversity by limiting our selection to a specified number of audiobooks. Moreover, we test the effect of batching short utterances vs long utterances on the downstream performance. We investigate two novel techniques for unsupervised in-domain data selection: pre-training loss based data selection (PL-based) and perplexity of byte pair encoded clustered units (PBPE). We enumerate all of our data selection criteria in Table 3.1. Appendix A has more details and statistics for some of the properties of the fine-tuning subsets associated with each data selection criterion.

Criterion	Description
PUR_RND	Sample 10 hours randomly
GNDR_DIV	Sampled subset contains 24 speakers of one gender (male / female)
UTTLN_DIV_RND_LNG_DUR_TAIL	Sample from the 15% of the utterances with the longest duration
UTTLN_DIV_RND_SHRT_DUR_TAIL	Sample from the 15% of the utterances with the shortest duration
UTTLN_DIV_RND_MIDDLE_DUR	Sample from the middle 15% of the utterances in terms of duration
SPK_DIV_RND	Sample utterances from a speci ed number of speakers (24 - 96)
BK_DIV_RND	Sample utterances from a speci ed number of books (16 - 64)
PRETRAIN_U_LOSS_AVG_NO_MASK	Compute the average unmasked pre-training loss for each utterance after turning o the mask. Sample from the 15% with the lowest loss (HEAD) vs the 15% with the highest loss (TAIL).
PRETRAIN_M_LOSS_AVG	Compute the average masked pre-training loss for each utterance. Sample from the 15% with the lowest loss (HEAD) vs the 15% with the highest loss (TAIL).
PERPLEXITY_5k_LM_15	Use PBPE to compute utterance score. Sample from the 15% with the lowest score (HEAD) vs the 15% with the highest score (TAIL).
PERPLEXITY_5k_LM_40_MIDDLE	Sample utterances with PBPE from the middle 40% of the data

Table 3.1: Description of data selection criteria for the fine-tuning subsets

3.1 Pre-training Loss Based Data Selection

We base our experiments on the HuBERT model and investigate the use of the HuBERT pre-training loss function as a means of data selection. Similar to wav2vec 2.0, HuBERT selects $p\%$ of the time steps as start indices for masking, and then spans of l time steps are masked. Using targets derived via K-means clustering of MFCCs or features extracted using a previous snapshot of the model, the cross-entropy loss is then computed over the masked and the unmasked time steps as L_m and L_u and the weighted sum of both is taken as

the final loss as defined in the following equation:

$$L = \quad L_m + (1 \qquad) L_u \tag{3.1}$$

For our data selection, we compute the score of each utterance as a function of the pre-training loss. First, we compute the utterance score as the average pre-training loss over all the masked frames in the utterance. We sort the utterances ascendingly based on the computed scores and select the first X hours or the last X hours, where X is either 1 or 10 in our experiments. Since the masking is inherently random, we hypothesize this may lead to a criterion that is close to pure random selection as the same utterance will get some different score each time we compute the loss. In light of this, we also consider turning off the mask and computing the utterance score as the average cross-entropy loss over all the unmasked frames. The advantage of this second method is that we get a deterministic score for each utterance. We sort the utterances ascendingly based on the computed scores and randomly sample utterances from both the top and the bottom 15% of the data. Figure 3.1 shows that the average unmasked loss per utterance spans a small range. A similar observation holds for the average masked loss per utterance as shown in figure 3.2.

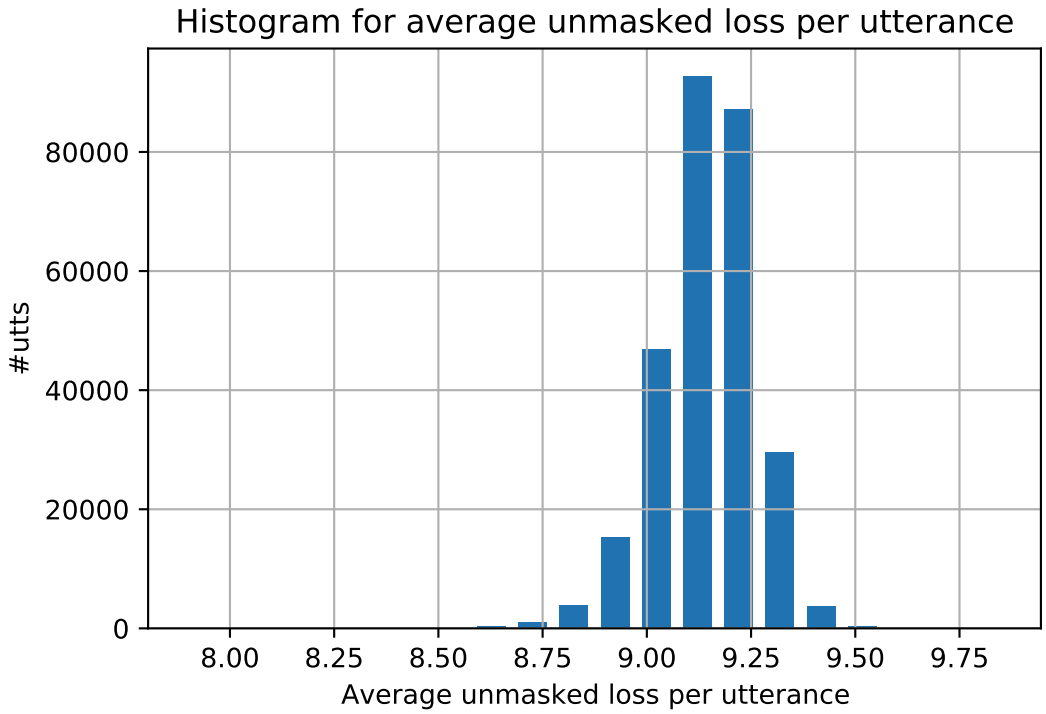

Figure 3.1: Histogram for the average unmasked loss per utterance in Librispeech

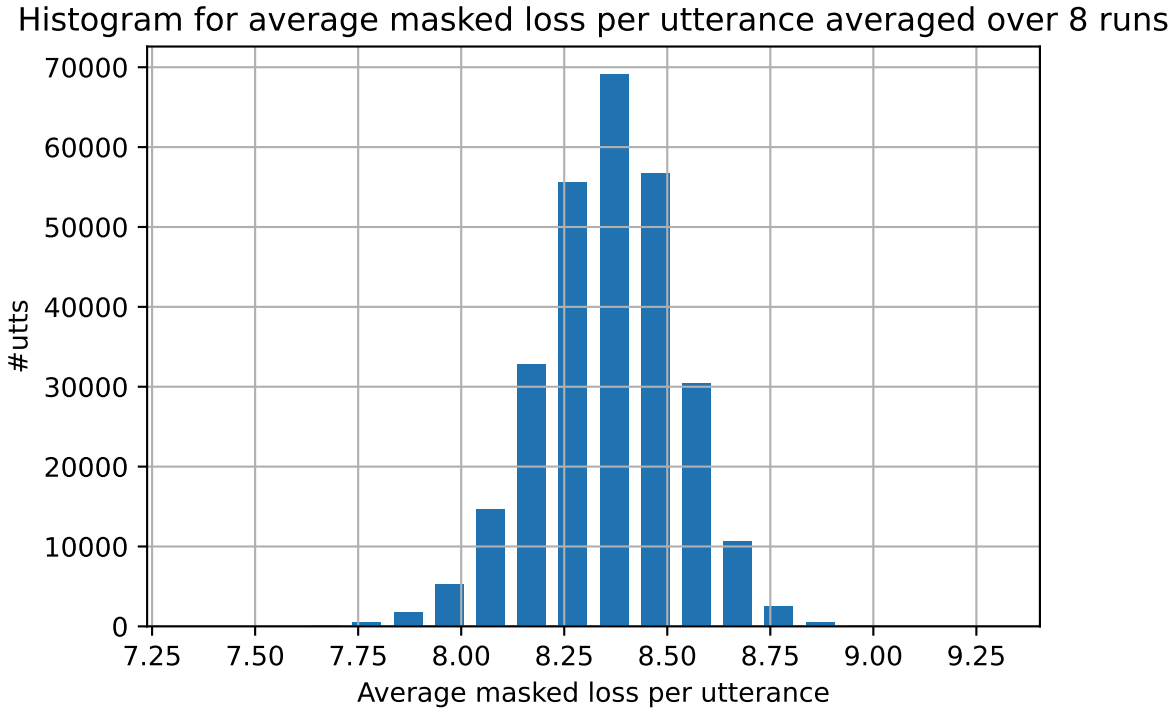

Figure 3.2: Histogram for the average masked loss per utterance in Librispeech. The masked loss per utterance is the mean value computed over 8 runs.

3.2 Perplexity of Byte Pair Encoded Clustered Units (PBPE)

HuBERT uses K-means clustering to generate noisy labels for the pre-training step. The primary reason this approach works is that even though the labels are noisy, they tend to be consistent [19]. We hypothesize that these labels are highly correlated to the tokens in the actual text transcripts in a way that enables us to utilize the same algorithms used for labeled data selection. First, we apply run-length encoding to collapse consecutive repetitions of the

frame-level HuBERT cluster labels. Next, we use sentencepiece [27] to train a BPE model on the run-length encoded sequences with a vocab size of 5k. Finally, we tokenize each sequence using that BPE model. We use fairseq [33] to train a 1-Layer LSTM language model with 512 hidden units over these unit-BPE sequences. We use the language model to compute the perplexity of each utterance in Librispeech train set. We then sort the utterances based on their perplexity and sample utterances from the top 15% and the bottom 15% of the whole train set to compare both criteria. Figure 3.3 shows the histogram for the utterance length in terms of number of BPE tokens, and figure 3.4 illustrates the perplexity range for the utterances in the training set.

In this chapter, we demonstrated the different methods that we used for selecting data for fine-tuning the HuBERT model. Moreover, we introduced two novel criteria for unsupervised data selection which are pre-training loss based data selection and PBPE. In the next chapter, we describe the setup for our experiments and analyze the results that we obtained.

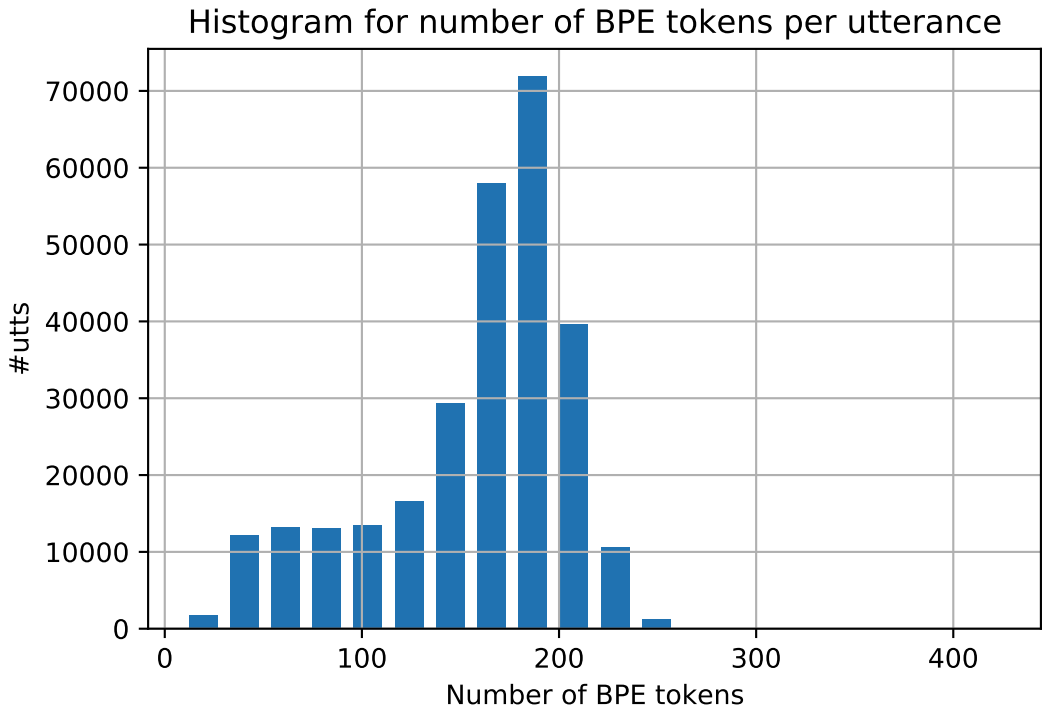

Figure 3.3: Histogram for the number of BPE tokens per utterance in Librispeech after tokenizing using a BPE model with a vocabulary size of 5k.

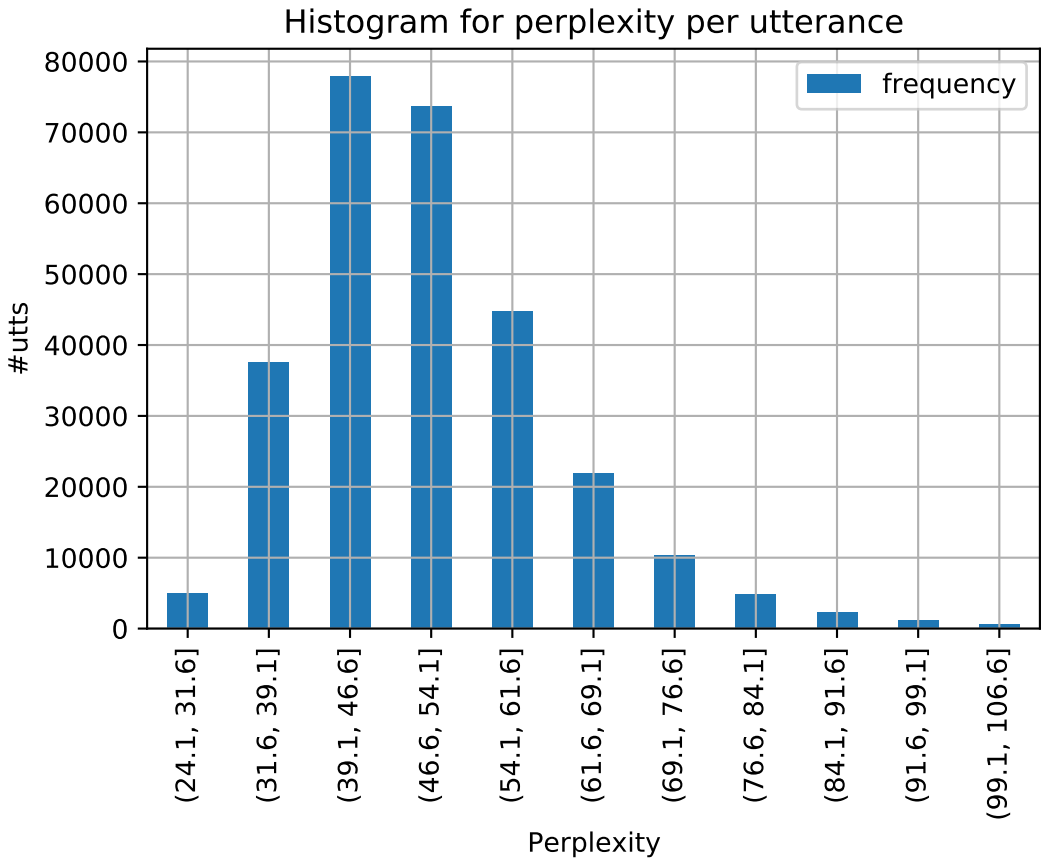

Figure 3.4: Histogram for the perplexity over BPE units for utterances in Librispeech. The histogram bins with fewer than 500 utterances are dropped for clarity.

Chapter 4: Experimental Setup and Results

In this chapter, we describe the setup used for our experiments in terms of the model and the data set used for training and evaluation. We demonstrate the results of the conducted experiments and provide a detailed analysis of the obtained results.

4.1 Model and Data

In our experiments, we use the HuBERT base model which is pre-trained on the full 960 hours of Librispeech [34] and is available on fairseq. Moreover, we use the same K-means clustering model with 500 clusters that is trained on the latent representations of HuBERT's 9th transformer layer after pre-training for 2 iterations.

For our data selection experiments, we use the full 960 hours of Librispeech as our data selection pool. Because each of our data selection criteria still utilizes random sampling in some form, we prepare 8 fine-tuning subsets, of 10 hours each. We experiment with different selection criteria for these subsets to probe how the model would behave under a differing number of speakers, a differing number of topics (books), and speaker gender bias. We also examine the impact of grouping utterances with similar lengths to see whether this would help the model learn better given the same number of up-

dates and the same maximum number of tokens per batch. In addition to this, we experiment with our proposed novel criteria: pre-training loss based data selection and PBPE. Table 3.1 summarizes the different data selection criteria that we used in our fine-tuning experiments. To examine whether we get the same observations at a more limited transcription budget, we experiment with selecting 1-hour subsets in a purely random fashion, as well as using our proposed novel selection criteria. We use Librispeech dev-other for validation and we test our models on both Librispeech test-clean and test-other.

4.2 Training

For the 10-hour experiments, we fine-tune the pre-trained HuBERT base model with target letter labels using each of the selected subsets described in table 3.1 for 25k updates. Similar to [19], we fix the convolutional waveform encoder for the whole training. We freeze the transformer encoder for the first 10k updates and then allow it to train with the rest of the model for the remaining updates. We use the adam optimizer [25] with betas set to 0.9 and 0.98 for model optimization. We fine-tune using a two stage learning rate scheduler, where the model ramps up to a peak learning rate of 2e-5 for the first 8k updates and then decays for the remaining updates. We use 2 gpus, and set the batch size to a maximum of 3200000 frames per gpu, and allow padding to the length of the longest utterance per batch. We start validation after 10k updates and keep the best checkpoint on dev-other subset. For the 1-hour setup, the transformer encoder is frozen for the first 5k updates, and

the training uses a warm-up of 4k steps, until it reaches a peak learning rate of 2e-5, then decays for 9k updates.

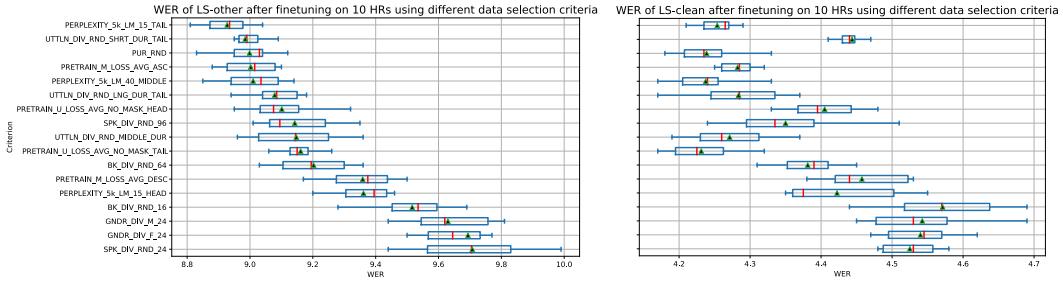

Figure 4.1: Box plot showing the WER on test-clean and test-other for different data selection criteria for 10-hour subsets. The green triangle represents the mean, while the red line represents the median. Also shown for each criterion are the minimum, maximum, 25th percentile, and 75th percentile.

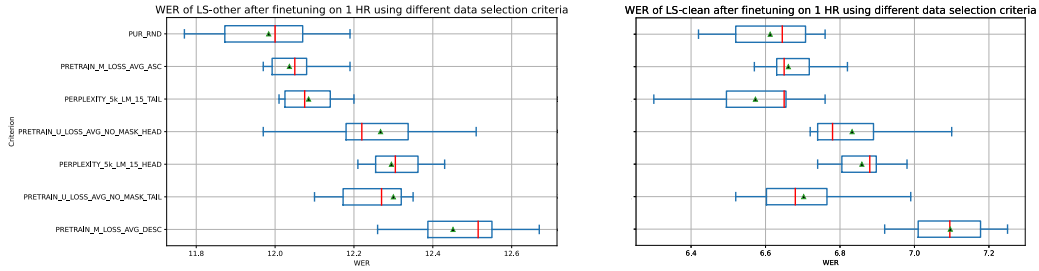

Figure 4.2: Box plot showing the WER on test-clean and test-other for different data selection criteria for 1-hour subsets. The green triangle represents the mean, while the red line represents the median. Also shown for each criterion are the minimum, maximum, 25th percentile, and 75th percentile.

4.3 Results

For each data selection criterion, we create 8 different random subsets and fine-tune the model on each of these subsets. We evaluate the best

checkpoint on dev-other for each fine-tuning subset and then record the mean, minimum, maximum, median, 25th percentile, and 75th percentile WER on both test-clean and test-other for each group of 8 random subsets. A 4-gram language model trained on Librispeech is used for all the decoding experiments. The box plots in Figure 4.1 summarize the results of our experiments using 10-hour fine-tuning subsets. Appendix B has a more detailed breakdown of the WER results of each experiment.

For PBPE, we see that for test-other, we score 8.93% WER on average when selecting from the 15% of utterances with the highest perplexity score (TAIL). On the other hand, we observe that the mean WER degrades to 9.36% when sampling from the 15% of the utterances with the lowest perplexity score (HEAD). We have a similar observation for test-clean, where the mean WER for TAIL is 4.25% compared to 4.42% for HEAD. For average masked pre-training loss, fine-tuning with the utterances with the smallest loss leads to a mean WER of 9.00% and 4.28% on test-other and test-clean respectively. However, using the largest loss utterances in fine-tuning leads to a mean WER of 9.36% on test-other and 4.46% on test-clean. For average unmasked pre-training loss, the mean WER on test-clean is 4.23% when sampling the fine-tuning subsets from the 15% utterances with the largest loss (TAIL), while it degrades to 4.41% when sampling from the 15% utterances with the smallest loss (HEAD). For test-other the mean WER is 9.1% for HEAD and 9.16% for TAIL, which is almost the same. We observe that the best results obtained from our proposed criteria are almost on par with the

pure random selection criterion which scores a mean WER of 9.00% and 4.24% on test-other and test-clean respectively. It is worthwhile mentioning that our experiments demonstrate that randomly selecting data from a diverse pool can lead to good performance when using this data for fine-tuning. In some real-world scenarios, it may be tempting to use whatever transcribed data is available for fine-tuning. However, selecting data from one audiobook with a single speaker, for example, may lead to sub-optimal results. On the other hand, randomly selecting data from a diverse pool makes the selected sample richer in vocabulary and more representative of different speakers and topics. Accordingly, it will most likely help to sample randomly from the pre-training data for transcription. In case the data selection pool is not diverse, it may be the case that our proposed selection criteria behave better than pure random selection.

Moreover, our experiments emphasize the importance of speaker diversity when selecting data for fine-tuning. We see that increasing the number of speakers from 24 to 96 leads to a significant boost in the mean WER reduction. It is interesting to see that the mean WER does not vary much regardless of whether the 24 speakers were of the same gender or selected randomly which suggests that the learned representations from SSL may be somewhat tolerant of gender bias in the fine-tuning data. However, we do find that topic diversity is crucial as it enriches the vocabulary that is present in the audio. Accordingly, the mean WER significantly improves when increasing the number of different audiobooks sampled for fine-tuning from 16 to 64. Lastly, when sam-

pling from the shortest 15% of the utterances, the mean WER on test-other significantly improves, but the opposite happens on test-clean.

When evaluating the 1-hour fine-tuning experiments, our observations are consistent with the 10-hour setup.

- PBPE performs better when selecting from the 15% of the utterances with the highest score (TAIL), compared to selecting from the utterances with the lowest perplexity score (HEAD). The mean WER for TAIL is 12.09% on test-other and 6.57% on test-clean, while the mean WER for HEAD is 12.3% on test-other and 6.86% on test-clean.

- The mean WER when selecting utterances with the lowest average masked pre-training loss (ASC) is better than when selecting utterances with the highest loss (DESC). The mean WER for the former is 12.04% and 6.66% on test-other and test-clean respectively, while the latter has a significantly worse mean WER of 12.45% on test-other and 7.1% on test-clean.

- The mean WER on test-clean when selecting utterances with the highest average unmasked pre-training loss (TAIL) is better than when selecting utterances with the lowest average unmasked pre-training loss (HEAD). It is 6.7% in the former compared to 6.83% in the latter case. However, the mean WER on test-other is almost on par when selecting the utterances with the lowest or the highest average unmasked pre-training loss, where it is 12.27% for the former and 12.3% for the latter.

- The best results obtained from our proposed selection criteria are almost on par with pure random data selection which has a mean WER of 11.98% and 6.61% respectively.

The box plot in figure 4.2 summarizes the results for our 1-hour fine-tuning experiments. In the next section, we dig deeper to analyze the results obtained on both the 1-hour and the 10-hour setups.

4.4 Analysis

We investigate how the different properties of the selected subsets correlate with each other and with the WER. Figures 4.3 and 4.4 show some interesting relations between the underlying properties of the selected subsets and the WER observed on both test-clean and test-other for the 10-hour and the 1-hour setups respectively. We observe a strong negative correlation between the WER on both test sets and the number of unique vocabulary words in the fine-tuning subset. This correlation is even stronger than the correlation between the WER and the total number of vocabulary words in these subsets. It is obvious that a strong positive correlation exists between the number of unique vocabulary words observed in the fine-tuning set and the perplexity computed over the BPE clustered units (PBPE). To compute this correlation, we averaged the perplexity over the total number of utterances in each fine-tuning subset (avg_ppl). Figure 4.5 highlights this correlation for the 10-hour fine-tuning subsets and we get the same observation for the 1-hour fine-tuning subsets as shown in figure 4.6. This correlation suggests that sampling from

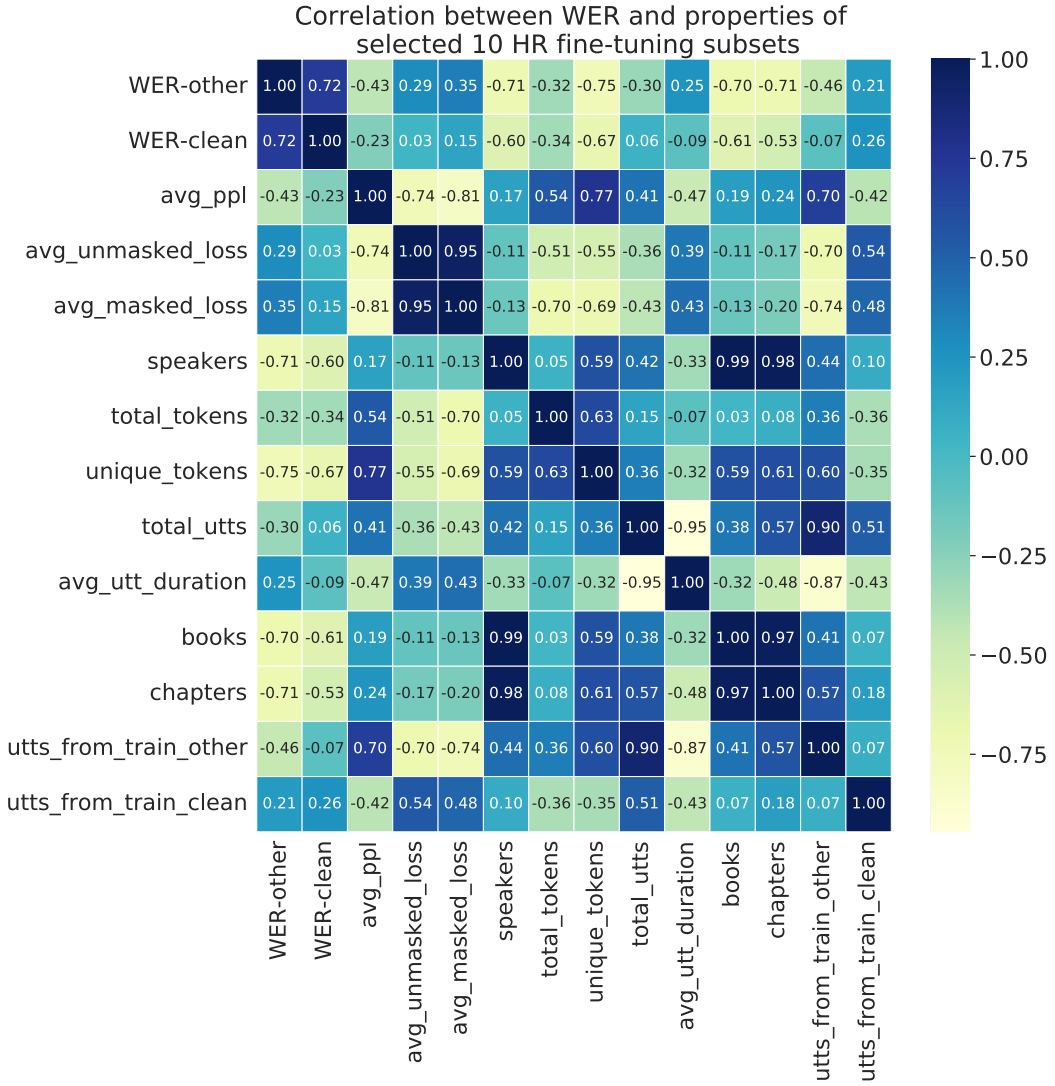

Figure 4.3: Correlation between the different properties of the selected 10-hour fine-tuning subsets and the WER on test-other and test-clean

the higher perplexity scoring utterances leads to more unique vocabulary words in the selected fine-tuning subset. This is an interesting observation because, in our unsupervised selection setup, we have no access to the transcription

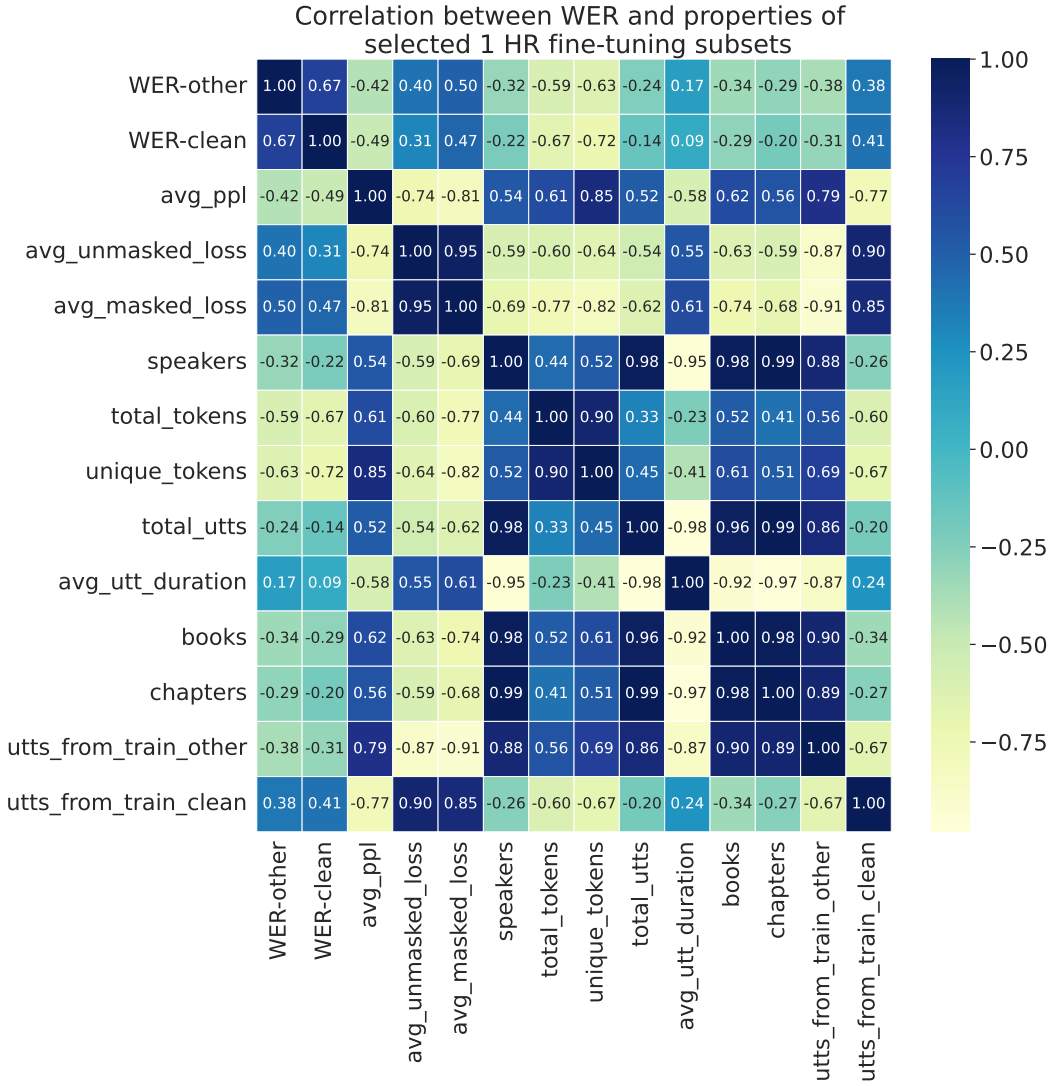

Figure 4.4: Correlation between the different properties of the selected 1-hour fine-tuning subsets and the WER on test-other and test-clean

tokens. However, we have access to the HuBERT clustered units that we can use as a proxy for the text. Figures 4.7 and 4.8 show the correlation between the WER on test-other and test-clean and the number of unique vocabulary

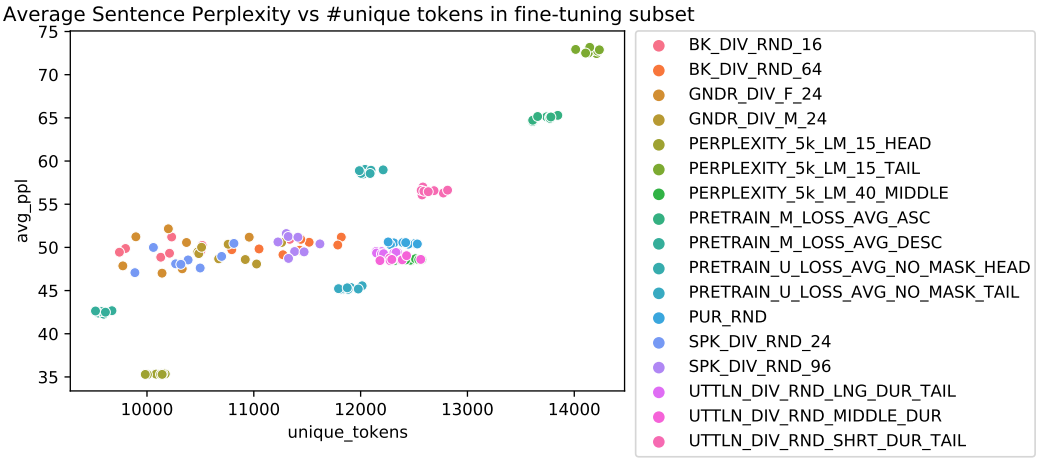

Figure 4.5: Correlation between the average sentence perplexity computed over the BPE clustered units and the number of unique vocabulary words appearing in the 10-hour fine-tuning subset

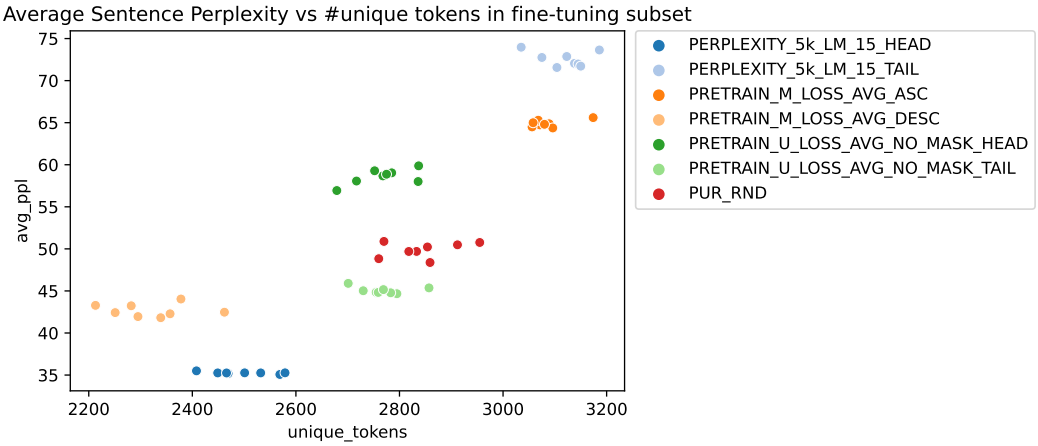

Figure 4.6: Correlation between the average sentence perplexity computed over the BPE clustered units and the number of unique vocabulary words appearing in the 1-hour fine-tuning subset

words in the fine-tuning subset and how the different selection criteria result in different numbers of unique vocabulary words in both the 10-hour and the

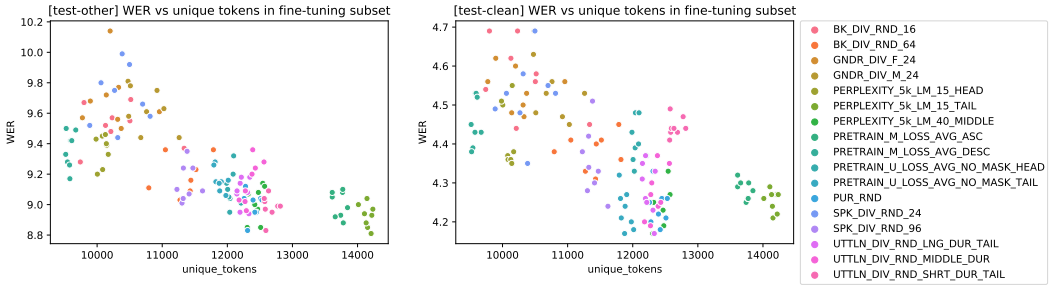

Figure 4.7: Correlation between WER on both test-other and test-clean and the number of unique vocabulary words appearing in the 10-hour fine-tuning subset

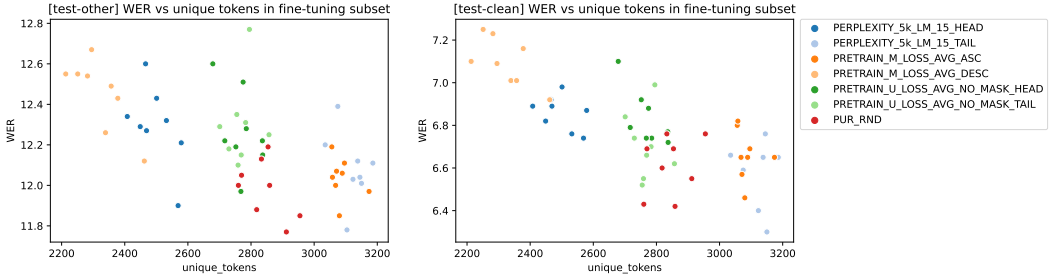

Figure 4.8: Correlation between WER on both test-other and test-clean and the number of unique vocabulary words appearing in the 1-hour fine-tuning subset

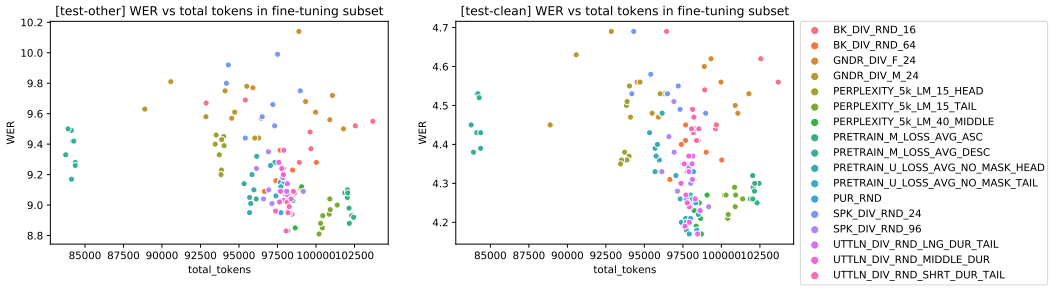

Figure 4.9: Correlation between WER on both test-other and test-clean and the total number of vocabulary words appearing in the 10-hour fine-tuning subset

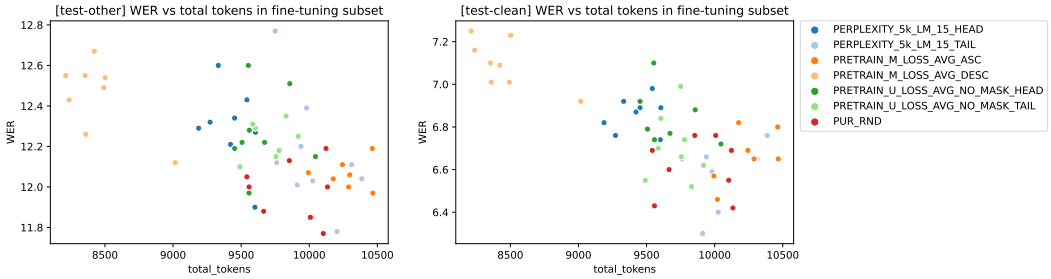

Figure 4.10: Correlation between WER on both test-other and test-clean and the total number of vocabulary words appearing in the 1-hour fine-tuning subset

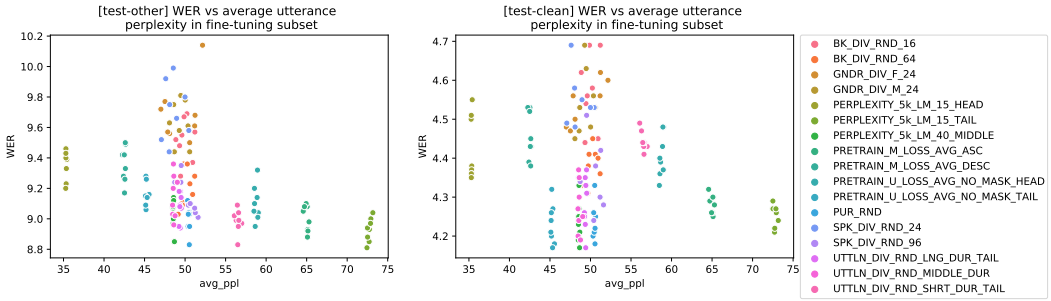

Figure 4.11: Correlation between WER on both test-other and test-clean and the average sentence perplexity over BPE clustered units in the 10-hour fine-tuning subset

1-hour setups. As pointed out, this correlation is stronger that the correlation between the WER on both test sets and the total number of vocabulary words in the fine-tuning subset, which is shown in figures 4.9 for the 10-hour subsets and 4.10 for the 1-hour subsets. Figures 4.11 and 4.12 illustrate the correlation between the WER on both test-other and test-clean and the average sentence perplexity over BPE clustered units and how the average sentence perplexity over BPE clustered units relates to the different data selection criteria in both the 10-hour and the 1-hour setups. Moreover, we observe a strong negative

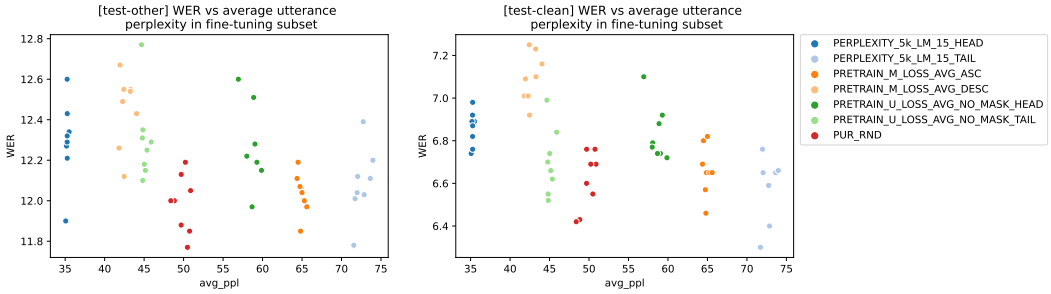

Figure 4.12: Correlation between WER on both test-other and test-clean and the average sentence perplexity over BPE clustered units in the 1-hour fine-tuning subset

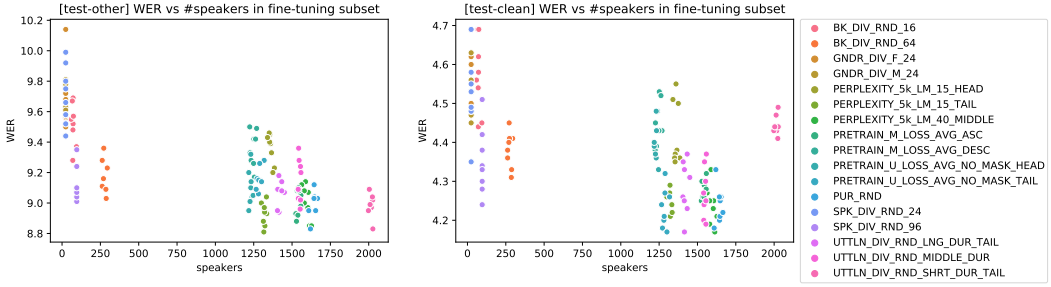

Figure 4.13: Correlation between WER on both test-other and test-clean and the total number of speakers in the 10-hour fine-tuning subset

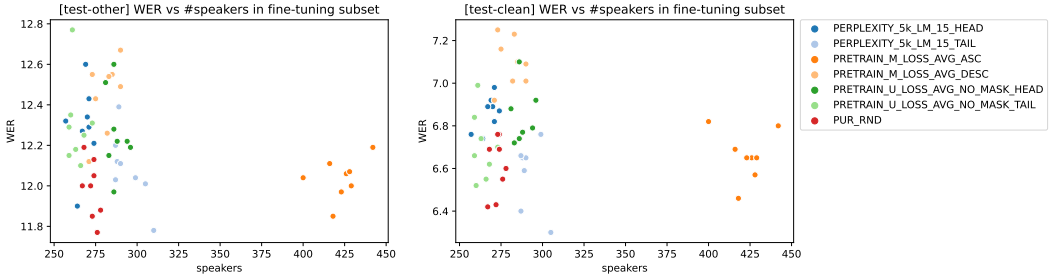

Figure 4.14: Correlation between WER on both test-other and test-clean and the total number of speakers in the 1-hour fine-tuning subset

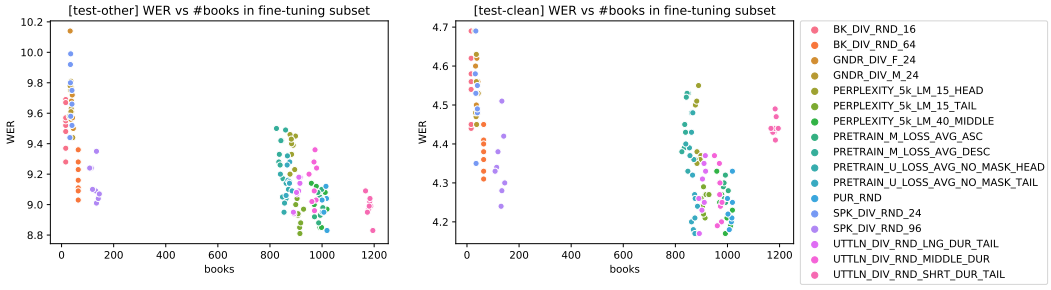

Figure 4.15: Correlation between WER on both test-other and test-clean and the total number of audiobooks in the 10-hour fine-tuning subset

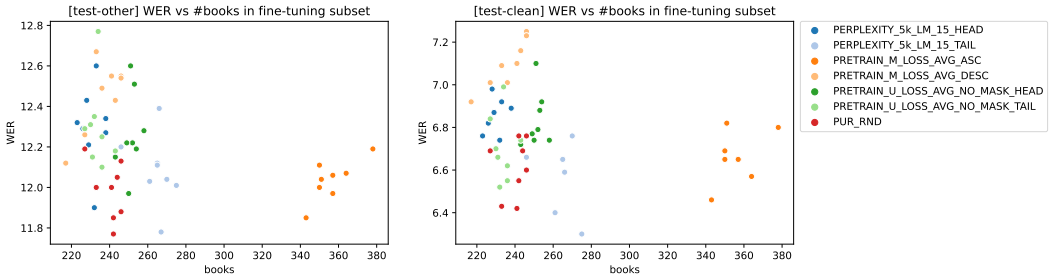

Figure 4.16: Correlation between WER on both test-other and test-clean and the total number of audiobooks in the 1-hour fine-tuning subset

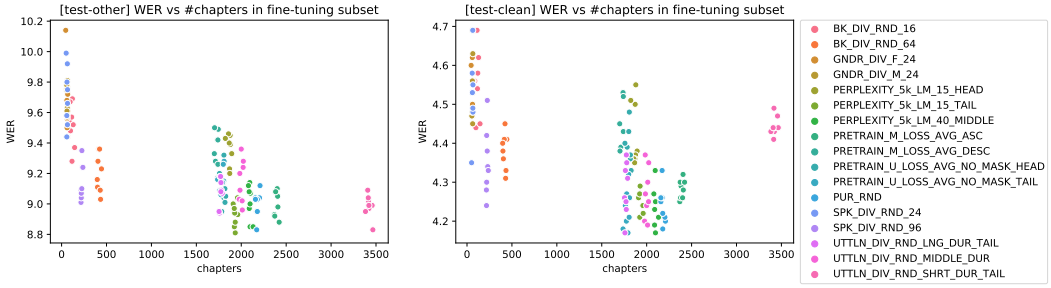

Figure 4.17: Correlation between WER on both test-other and test-clean and the total number of chapters in the 10-hour fine-tuning subset

correlation between the number of speakers and the WER and similarly for the number of audiobooks or chapters. Figures 4.13, 4.15 and 4.17 highlight

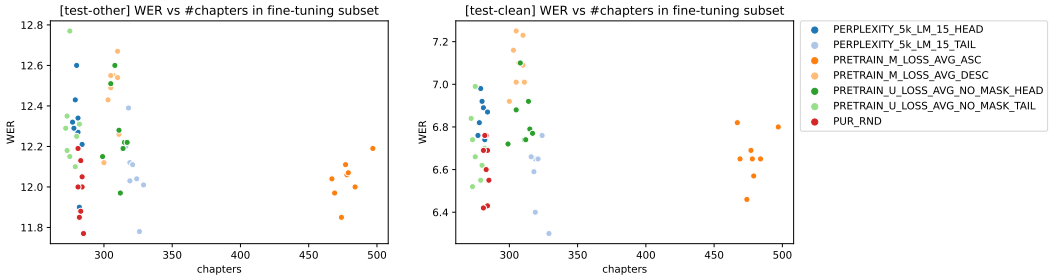

Figure 4.18: Correlation between WER on both test-other and test-clean and the total number of chapters in the 1-hour fine-tuning subset

these correlations for the 10-hour subsets, while figures 4.14, 4.16 and 4.18 demonstrate these correlations for the 1-hour subsets.

Driven by these observations, we conduct two experiments. In the first experiment, we select utterances from all the speakers in the top 15% highest scoring utterances in terms of perplexity over BPE clustered units. This ensures speaker diversity and maximizes the number of unique vocabulary words, and enables us to beat pure random selection on both test-clean and test-other. In the second experiment, we select utterances from almost every book in the top 15% highest scoring utterances in terms of perplexity. We arrive at similar results where we are able to beat pure random selection. Finally, we compare our results to fine-tuning HuBERT base model with libri-light 10-hour subset [22] and observe that our techniques are scoring better in terms of WER with a large margin. Tables 4.1 and 4.2 summarize the results of these experiments.

In Figures 4.3 and 4.4, we observe an interesting correlation between the average unmasked loss and the number of utterances selected from train-

other or train-clean. As the average unmasked loss decreases, sampling is more biased to the train-other subset, which can account for scoring a lower mean WER on test-other when sampling from the smaller loss utterances compared to the higher loss ones. However, as the average unmasked loss increases, sampling is more biased to train-clean subset, leading to an improved mean

Fine-tuning subset	WER-other	WERR over libri-light	WERR over PUR_RND
Libri-light	9.61	0.00%	-6.78%
PPL	8.93	**7.08%**	**0.78%**
PPL+speaker diversity	8.89	**7.49%**	**1.22%**
PPL+book diversity	8.8	**8.43%**	**2.22%**

Table 4.1: Librispeech test-other results' summary for PBPE experiments and fine-tuning with libri-light. In the table, PPL refers to PERPLEXITY_5k_LM_15_TAIL. The WER results for our criteria are the mean WER computed over 8 runs. WERR over libri-light refers to the word error rate reduction obtained when using each fine-tuning subset relative to fine-tuning with libri-light. WERR over PUR_RND refers to the word error rate reduction obtained when using each fine-tuning subset relative to fine-tuning with a randomly selected subset.

Fine-tuning subset	WER-clean	WERR over libri-light	WERR over PUR_RND
Libri-light	4.48	0.00%	-5.66%
PPL	4.25	**5.05%**	-0.32
PPL+speaker diversity	4.06	**9.38%**	**4.25%**
PPL+book diversity	4.21	**6.03%**	**0.71%**

Table 4.2: Librispeech test-clean results' summary for PBPE experiments and fine-tuning with libri-light. In the table, PPL refers to PERPLEXITY_5k_LM_15_TAIL. The WER results for our criteria are the mean WER computed over 8 runs. WERR over libri-light refers to the word error rate reduction obtained when using each fine-tuning subset relative to fine-tuning with libri-light. WERR over PUR_RND refers to the word error rate reduction obtained when using each fine-tuning subset relative to fine-tuning with a randomly selected subset.

WER on test-clean when sampling from the higher loss utterances. In case of the average masked loss, the mean WER on both test-other and test-clean is better when sampling from the lower loss utterances. Despite the correlation directions with the number of utterances from train-clean and train-other being maintained in case of average masked loss, the improved mean WER on test-clean when sampling from lower loss utterances suggests that other factors are involved. Our investigation shows that for average masked loss criterion, we end up with more speakers and more topics when sampling from the lower loss utterances compared to the higher loss utterances. However, the same does not happen for the average unmasked loss criterion which can account for the different behaviour. We hypothesize that since all the utterances in our selection pool are included in the pre-training stage of the model, the pre-training loss based criterion will be highly impacted by how frequently each utterance was fed to the model during pre-training as well as the whole pre-training setup causing some biases in the selection process.

In this chapter, we presented our experimental setup and the results obtained in the experiments conducted for fine-tuning data selection. Moreover, we analyzed the obtained results and shared various insights. In the next chapter, we summarize our conclusions and propose directions for future work.

Chapter 5: Conclusions and Future Work

In this thesis, we investigate different criteria for selecting data for fine-tuning a self-supervised speech model to perform ASR. We select in-domain data from a large pool of unlabeled data using unsupervised techniques. We build our study on top of HuBERT base model because it allows us to make use of the K-means clustered units as a proxy for the transcription. Our study aims at providing answers to the following questions:

- **Given an unlabeled pool of in-domain data, are we guaranteed optimal or near-optimal performance if we randomly select data for fine-tuning in low-resource scenarios?**

  Our experiments show that pure random data selection is a good technique that can be hard to beat in both the 1-hour and the 10-hour setups. The analysis conducted shows that selecting data in a purely random way guarantees that the selected samples exhibit a lot of diversity regarding speakers, topics, and vocabulary words as long as the data selection pool is inherently diverse. If the data selection pool is not sufficiently diverse, it is possible that our proposed novel criteria may work better than pure random data selection.

- **Can we arrive at an unsupervised technique for data selection from a pool of in-domain unlabeled data that guarantees**

optimal or near-optimal performance on the ASR task when fine-tuning a self-supervised speech model like HuBERT?

Our study investigates the possibility of arriving at unsupervised data selection techniques that show better performance than pure random data selection. We devise two novel techniques in that regard. In the first technique, we make use of the pre-training loss of the HuBERT model to compute a score for each utterance in the data pool. We experiment with the masked pre-training loss as well as the unmasked pre-training loss. We show that selecting utterances with the lowest masked pre-training loss leads to a much better performance than selecting the utterances with the highest masked pre-training loss. However, the analysis conducted shows that data selection based on pre-training loss can inherit some biases from the pre-training stage depending on how the examples are fed to the model during pre-training. This is demonstrated in the correlations between the number of utterances selected from train-clean or train-other and the pre-training loss.

The second technique we propose for unsupervised data selection is PBPE. It makes use of the labels that are used in pre-training the HuBERT model, which are generated by applying K-means clustering either on MFCC features or the latent representations extracted from some intermediate layer in the transformer stack of the HuBERT model. We treat these labels as a proxy for text, apply BPE on top of them, and train a language model on the resulting pseudo-transcripts. The

language model is used to compute the perplexity score for each utterance. Our experiments show that PBPE works very well on the ASR task and performs better or on par with pure random data selection in the different setups when selecting the utterances with the highest perplexity score. PBPE beats pure random data selection when enforcing speaker diversity by selecting utterances from all the speakers in the top 15% highest scoring utterances in terms of perplexity. The analysis conducted shows that the average sentence perplexity over BPE clustered units in the fine-tuning subsets positively correlates with the number of unique vocabulary words in the fine-tuning subsets. Since our analysis demonstrates that the WER gets better as the fine-tuning subset has more unique vocabulary words, PBPE can be used to drive more vocabulary words into the fine-tuning subset in the absence of transcriptions, and hence provides better performance on the ASR task.

- **Are the learned HuBERT representations robust enough in the sense that we can guarantee the same performance regardless of the number of speakers in the fine-tuning subsets as well as their genders?**

  Our investigation shows that speaker diversity in the fine-tuning subsets is important in achieving better WER. This has been demonstrated in various ways. Increasing the number of speakers in the fine-tuning subset from 24 to 96 leads to a boost in the WER reduction on both test-clean and test-other. In addition to this, PBPE achieves the best performance

when combined with enforcing speaker diversity in the selected subsets. Moreover, the data selection criteria that result in a diversity of speakers lead to much better results than those obtained by fine-tuning with libri-light which has 24 speakers (12 female speakers and 12 male speakers). It is interesting to observe that the average WER obtained at a fixed number of speakers is almost the same regardless of the choice of speakers' genders.

- **Does the ASR performance get better when enriching the vocabulary in the audio by representing more topics in the fine-tuning subsets?**

  Our experiments show that topic diversity in the fine-tuning subsets is important in multiple ways. Increasing the number of books in the selected fine-tuning subsets from 16 to 64 results in improving the WER significantly. In addition to this, our analysis shows a negative correlation between the number of audio books in the fine-tuning subsets and the WER. A similar observation holds for the relation between the number of chapters and the WER. Intuitively, adding more topics in the fine-tuning subsets enriches the vocabulary that is present in these subsets and hence improves the WER, which aligns with our observation that a correlation exists between the number of unique vocabulary words and the WER.

- **When the fine-tuning subset is in the range of 1 to 10 hours, do the WER results exhibit high variance based on the utterances**

selected for each criterion?

It is obvious that working in a low resource setup, in which the fine-tuning data subsets can only be a few hours (1 or 10 hours in our experiments) is associated with the risk of having high variance in the ASR performance depending on the choice of utterances in each fine-tuning subset. For this reason, we prepare 8 fine-tuning subsets for each data selection criterion to test this behavior. Our experiments show that the standard deviation of the WER associated with each data selection criterion in the 10-hour setup is small (0.06 on average for the WER on test-clean and 0.12 on average for the WER on test-other). A similar observation holds for the 1-hour setup, where the standard deviation of the WER is 0.13 on test-clean and 0.17 on test-other on average. Accordingly, we can draw distinctions between the data selection criteria and sort them based on the downstream performance on the ASR task.

Several questions still remain open for future work. We would like to investigate whether our observations and conclusions hold when the target domain differs from the pre-training domain. In addition to this, it would be interesting to evaluate our proposed approaches on test sets other than Librispeech. Furthermore, we would like to investigate perplexity-based selection for a target domain when the unlabeled data pool has a diverse set of domains.

Appendices

Appendix A: Exact Statistics for the Selected Subsets for Fine-tuning Experiments

A.1 Statistics for the 10-hour Subsets

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	books	chapters	utterances from train-other	utterances from train-clean
BK_DIV_RND_16_0	50.23	9.13	39	33	72	95409	10517	2996	31.85	80	2	12.03	22.80	1.43	16	120	1379	1617
BK_DIV_RND_16_1	48.86	9.15	48	24	72	102554	10129	2914	35.19	70	2	12.44	17.34	1.71	16	129	1718	1196
BK_DIV_RND_16_2	49.32	9.15	33	38	71	99619	10211	2935	33.94	64	2	12.23	19.65	1.63	16	99	1382	1553
BK_DIV_RND_16_3	49.87	9.13	34	33	67	92867	9799	2866	32.40	80	1	12.53	22.80	0.83	16	95	1671	1195
BK_DIV_RND_16_4	49.44	9.13	43	27	70	98916	9742	2847	34.74	69	2	12.65	18.09	1.49	16	116	1347	1500
BK_DIV_RND_16_5	49.61	9.14	35	24	59	103692	10504	2922	35.49	75	1	12.32	22.91	1.26	16	87	1201	1721
BK_DIV_RND_16_6	51.21	9.13	37	37	74	96439	10231	2978	32.38	67	1	12.07	17.17	1.34	16	112	1798	1180
BK_DIV_RND_16_7	50.93	9.14	52	42	94	99671	11337	2898	34.39	72	1	12.39	17.34	1.41	16	145	1367	1531
BK_DIV_RND_64_0	51.19	9.15	135	127	262	100022	11819	2952	33.88	66	2	12.25	17.15	1.58	64	405	1628	1324
BK_DIV_RND_64_1	49.73	9.14	150	112	262	99011	10794	2886	34.31	72	1	12.50	17.77	1.28	64	400	1535	1351
BK_DIV_RND_64_2	49.15	9.16	155	133	288	97783	11273	2897	33.75	74	2	12.46	22.59	1.43	64	434	1539	1358
BK_DIV_RND_64_3	50.59	9.14	149	146	295	98482	11518	2927	33.65	69	2	12.35	19.52	1.44	64	444	1527	1400
BK_DIV_RND_64_4	49.82	9.15	137	137	274	97667	11049	2933	33.30	65	2	12.29	17.32	1.66	64	415	1490	1443
BK_DIV_RND_64_5	50.28	9.15	127	145	272	97676	11786	2892	33.77	68	2	12.45	17.28	1.81	64	423	1606	1286
BK_DIV_RND_64_6	50.93	9.14	154	116	270	97396	11440	2979	32.69	63	1	12.07	17.13	1.24	64	397	1663	1316
BK_DIV_RND_64_7	49.68	9.15	155	131	286	96654	11429	2897	33.36	63	2	12.37	19.65	1.58	64	430	1489	1408

Table A.1: Statistics for the 10-hour fine-tuning subsets selected to test the impact of book diversity

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	books	chapters	utterances from train-other	utterances from train-clean
GNDR_DIV_F_24_0	51.23	9.13	24	0	24	99326	9896	2943	33.75	65	2	12.24	17.21	1.81	38	59	1689	1254
GNDR_DIV_F_24_2	52.15	9.12	24	0	24	98889	10200	2956	33.45	62	1	12.19	17.19	1.13	33	46	1512	1444
GNDR_DIV_F_24_4	50.57	9.14	24	0	24	101785	10370	2940	34.62	86	2	12.25	26.21	1.80	45	61	1504	1436
GNDR_DIV_F_24_5	47.01	9.17	24	0	24	101068	10141	2907	34.77	67	1	12.38	18.78	1.35	41	68	787	2120
GNDR_DIV_F_24_6	48.10	9.13	24	0	24	100899	10316	2871	35.14	68	1	12.54	19.30	1.51	36	61	1105	1766
GNDR_DIV_F_24_7	51.18	9.12	24	0	24	99973	10957	3053	32.75	74	1	11.80	18.81	1.43	39	69	1455	1598
GNDR_DIV_F_24_8	47.52	9.16	24	0	24	95905	10330	2942	32.60	65	2	12.23	17.31	1.49	37	61	519	2423
GNDR_DIV_F_24_9	47.86	9.16	24	0	24	94528	9774	2949	32.05	66	1	12.20	17.28	1.27	42	61	1144	1805
GNDR_DIV_M_24_0	48.08	9.17	0	24	24	88885	11026	2950	30.13	59	1	12.21	17.21	1.08	37	63	1104	1846
GNDR_DIV_M_24_1	48.68	9.17	0	24	24	96284	10671	2872	33.53	64	3	12.55	19.64	2.03	36	59	1338	1534
GNDR_DIV_M_24_2	50.37	9.16	0	24	24	94720	10760	2923	32.41	63	1	12.32	16.94	1.24	36	60	914	2009
GNDR_DIV_M_24_3	50.57	9.15	0	24	24	96026	11260	2941	32.65	63	1	12.26	20.00	1.08	43	66	1738	1203
GNDR_DIV_M_24_4	49.47	9.15	0	24	24	90572	10475	2997	30.22	60	1	12.02	19.64	1.24	36	65	989	2008
GNDR_DIV_M_24_5	48.60	9.17	0	24	24	94096	10920	2958	31.81	64	1	12.17	17.15	1.24	35	54	1048	1910
GNDR_DIV_M_24_6	49.28	9.15	0	24	24	92854	10485	2922	31.78	62	2	12.33	17.26	1.36	30	59	1265	1657
GNDR_DIV_M_24_7	50.01	9.15	0	24	24	95508	10511	2855	33.45	62	3	12.62	17.06	1.86	33	58	1220	1635

Table A.2: Statistics for the 10-hour fine-tuning subsets selected to probe gender bias

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
PERPLEXITY_5k_LM_15_HEAD_0	35.35	9.20	670	696	1366	93719	10171	2940	31.88	66	2	12.26	20.09	1.51	883	1894	980	1960
PERPLEXITY_5k_LM_15_HEAD_10	35.33	9.19	686	669	1355	94006	10080	2941	31.96	63	1	12.25	17.09	1.51	898	1876	972	1969
PERPLEXITY_5k_LM_15_HEAD_13	35.41	9.19	661	699	1360	94033	10149	2954	31.83	66	1	12.20	18.35	1.61	889	1879	972	1982
PERPLEXITY_5k_LM_15_HEAD_14	35.33	9.19	690	695	1385	93862	10085	2945	31.87	61	3	12.23	17.42	1.73	895	1870	1016	1929
PERPLEXITY_5k_LM_15_HEAD_3	35.27	9.19	682	693	1375	93834	10004	2931	32.01	66	2	12.29	18.35	1.57	877	1872	970	1961
PERPLEXITY_5k_LM_15_HEAD_4	35.30	9.19	671	680	1351	93516	10127	2927	31.95	67	2	12.31	17.27	1.66	877	1858	995	1932
PERPLEXITY_5k_LM_15_HEAD_7	35.30	9.20	667	687	1354	93453	10142	2967	31.50	65	2	12.15	17.28	1.54	883	1866	989	1978
PERPLEXITY_5k_LM_15_HEAD_9	35.29	9.20	664	676	1340	93879	9984	2917	32.18	67	1	12.35	17.24	1.17	882	1823	948	1969
PERPLEXITY_5k_LM_15_TAIL_0	72.44	9.07	604	712	1316	100199	14208	3475	28.83	70	1	10.37	22.19	1.19	915	1933	2564	911
PERPLEXITY_5k_LM_15_TAIL_2	72.92	9.06	591	709	1300	101380	14013	3459	29.31	85	1	10.42	23.70	1.23	899	1910	2582	877
PERPLEXITY_5k_LM_15_TAIL_3	72.71	9.07	620	703	1323	100413	14154	3462	29.00	71	1	10.41	19.36	1.30	914	1923	2559	903
PERPLEXITY_5k_LM_15_TAIL_4	72.52	9.07	615	698	1313	100301	14132	3453	29.05	66	1	10.43	19.12	1.07	918	1933	2525	928
PERPLEXITY_5k_LM_15_TAIL_5	72.74	9.07	640	695	1335	100440	14219	3411	29.45	80	1	10.56	22.80	0.92	908	1961	2516	895
PERPLEXITY_5k_LM_15_TAIL_7	73.17	9.06	606	727	1333	100896	14145	3453	29.22	71	1	10.43	22.19	0.92	904	1958	2589	864
PERPLEXITY_5k_LM_15_TAIL_8	72.51	9.07	616	705	1321	100857	14106	3458	29.17	76	1	10.42	19.63	1.11	908	1927	2545	913
PERPLEXITY_5k_LM_15_TAIL_9	72.88	9.07	625	694	1319	100923	14235	3466	29.12	72	1	10.40	18.61	1.07	929	1922	2614	852
PERPLEXITY_5k_LM_40_MIDDLE_0	48.50	9.15	758	848	1606	98564	12323	2819	34.96	64	1	12.78	17.20	1.32	990	2101	1447	1372
PERPLEXITY_5k_LM_40_MIDDLE_1	48.65	9.15	780	833	1613	98702	12306	2814	35.08	67	3	12.81	17.13	1.93	992	2098	1438	1376
PERPLEXITY_5k_LM_40_MIDDLE_2	48.49	9.15	766	814	1580	98364	12477	2790	35.26	64	2	12.92	17.00	1.94	1012	2088	1395	1395
PERPLEXITY_5k_LM_40_MIDDLE_3	48.52	9.15	773	830	1603	98233	12466	2832	34.69	69	2	12.72	17.58	1.38	1019	2089	1428	1404
PERPLEXITY_5k_LM_40_MIDDLE_4	48.68	9.15	780	847	1627	98646	12512	2847	34.65	71	1	12.65	21.11	1.74	998	2132	1471	1376
PERPLEXITY_5k_LM_40_MIDDLE_5	48.55	9.15	741	839	1580	98166	12422	2831	34.68	69	2	12.73	20.03	1.85	970	2098	1478	1353
PERPLEXITY_5k_LM_40_MIDDLE_6	48.58	9.15	754	834	1588	98054	12544	2843	34.49	64	1	12.67	20.03	1.40	959	2095	1468	1375
PERPLEXITY_5k_LM_40_MIDDLE_7	48.59	9.15	748	815	1563	99068	12573	2831	34.99	69	2	12.73	17.25	1.86	978	2080	1476	1355

Table A.3: Statistics for the 10-hour fine-tuning subsets selected based on perplexity of Byte Pair Encoded (BPE) clustered units

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
PRETRAIN_M_LOSS_AVG_ASC_0	65.13	8.98	753	772	1525	102316	13741	5625	18.19	62	1	6.40	19.87	0.83	993	2369	4362	1263
PRETRAIN_M_LOSS_AVG_ASC_1	65.07	8.98	738	796	1534	101878	13769	5689	17.91	67	1	6.33	21.64	1.04	988	2378	4400	1289
PRETRAIN_M_LOSS_AVG_ASC_2	64.58	8.98	760	798	1558	102031	13611	5707	17.88	68	1	6.31	19.87	0.83	992	2405	4357	1350
PRETRAIN_M_LOSS_AVG_ASC_3	64.94	8.98	744	798	1542	102038	13773	5704	17.89	68	1	6.32	17.13	0.92	999	2400	4418	1286
PRETRAIN_M_LOSS_AVG_ASC_4	64.73	8.98	737	807	1544	102064	13611	5639	18.10	69	1	6.39	17.07	0.92	984	2383	4358	1281
PRETRAIN_M_LOSS_AVG_ASC_5	65.29	8.97	764	793	1557	102147	13846	5685	17.97	65	1	6.34	17.07	0.92	1003	2408	4379	1306
PRETRAIN_M_LOSS_AVG_ASC_6	65.16	8.98	743	795	1538	102446	13657	5609	18.26	67	1	6.42	19.19	0.92	971	2373	4316	1293
PRETRAIN_M_LOSS_AVG_ASC_7	65.10	8.97	753	776	1529	102162	13781	5643	18.10	68	1	6.39	17.22	1.04	987	2422	4306	1337
PRETRAIN_M_LOSS_AVG_DESC_0	42.40	9.29	572	661	1233	84369	9545	3517	23.99	54	1	10.25	16.98	1.07	834	1713	1116	2401
PRETRAIN_M_LOSS_AVG_DESC_1	42.53	9.29	565	689	1254	84128	9582	3547	23.72	56	1	10.16	17.06	1.07	850	1737	1194	2353
PRETRAIN_M_LOSS_AVG_DESC_2	42.27	9.29	573	678	1251	84191	9595	3492	24.11	62	1	10.32	17.13	1.13	845	1737	1165	2327
PRETRAIN_M_LOSS_AVG_DESC_3	42.67	9.29	586	683	1269	84118	9672	3557	23.65	58	1	10.13	17.10	1.07	860	1744	1204	2353
PRETRAIN_M_LOSS_AVG_DESC_4	42.58	9.29	572	677	1249	84370	9577	3539	23.84	54	1	10.18	17.22	1.07	839	1740	1149	2390
PRETRAIN_M_LOSS_AVG_DESC_5	42.62	9.29	556	668	1224	83903	9525	3530	23.77	60	1	10.21	17.22	1.13	825	1703	1178	2352
PRETRAIN_M_LOSS_AVG_DESC_6	42.50	9.29	581	681	1262	84276	9611	3480	24.22	54	1	10.35	17.03	1.07	842	1740	1195	2285
PRETRAIN_M_LOSS_AVG_DESC_7	42.64	9.29	562	664	1226	83752	9518	3503	23.91	57	1	10.28	17.22	1.07	836	1700	1186	2317
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_0	58.65	8.96	631	587	1218	95689	12037	3365	28.44	71	1	10.71	19.12	1.34	854	1783	2763	602
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_1	59.03	8.96	663	576	1239	96119	12041	3402	28.25	62	1	10.59	21.64	1.05	862	1806	2790	612
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_2	58.50	8.96	656	590	1246	95692	12024	3353	28.54	72	1	10.74	17.27	1.24	848	1827	2731	622
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_3	58.93	8.96	641	590	1231	96154	12097	3364	28.58	68	1	10.71	17.24	1.32	867	1807	2797	567
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_4	58.57	8.96	661	570	1231	95967	12005	3352	28.63	67	1	10.75	21.01	0.83	871	1818	2759	593
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_5	58.56	8.96	620	598	1218	95847	12089	3318	28.89	66	1	10.86	17.14	1.07	840	1758	2759	559
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_6	58.89	8.96	639	594	1233	95335	11988	3359	28.38	71	1	10.73	21.64	1.08	862	1801	2760	599
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_7	58.98	8.96	627	599	1226	95759	12211	3380	28.33	65	1	10.66	21.01	1.27	856	1819	2767	613
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_0	45.36	9.31	570	718	1288	97983	11907	2913	33.64	63	1	12.37	19.64	1.51	874	1778	825	2088
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_1	45.56	9.31	556	716	1272	97723	12014	2939	33.25	66	1	12.26	17.22	1.48	869	1737	836	2103
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_2	45.18	9.31	563	718	1281	97398	11978	2927	33.28	72	2	12.31	20.68	1.71	862	1774	819	2108
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_3	45.14	9.30	555	711	1266	97981	11883	2955	33.16	66	2	12.19	17.09	1.89	885	1762	785	2170
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_4	45.11	9.31	573	710	1283	97686	11824	2951	33.10	63	2	12.21	17.08	1.82	875	1805	804	2147
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_5	45.17	9.31	584	734	1318	97324	11812	2938	33.13	62	2	12.26	17.09	1.64	876	1809	859	2079
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_6	45.22	9.31	591	698	1289	97048	11793	2932	33.10	63	1	12.29	17.39	1.51	866	1800	837	2095
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_7	45.32	9.31	582	719	1301	97912	11875	2928	33.44	63	2	12.31	17.10	1.80	875	1792	795	2133

Table A.4: Statistics for the 10-hour fine-tuning subsets selected based on pre-training loss

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
PUR_RND_10	50.52	9.14	784	836	1620	98183	12311	2942	33.37	67	1	12.25	20.88	0.83	1019	2172	1602	1340
PUR_RND_11	50.31	9.14	788	856	1644	97718	12279	2932	33.33	68	1	12.29	18.09	1.24	1015	2209	1529	1403
PUR_RND_13	50.63	9.14	788	860	1648	98337	12258	2926	33.61	75	1	12.31	19.18	1.58	1019	2158	1568	1358
PUR_RND_14	50.46	9.14	790	857	1647	97999	12510	2951	33.21	73	1	12.21	21.01	1.53	1017	2203	1579	1372
PUR_RND_4	50.56	9.14	804	862	1666	98498	12395	2925	33.67	67	1	12.32	17.32	1.48	1002	2179	1574	1351
PUR_RND_6	50.37	9.14	805	849	1654	97722	12447	2920	33.47	68	1	12.34	19.88	1.43	1001	2173	1534	1386
PUR_RND_7	50.56	9.14	776	833	1609	98409	12422	2918	33.72	70	3	12.35	23.39	1.88	1007	2179	1543	1375
PUR_RND_9	50.40	9.14	782	861	1643	98121	12531	2954	33.22	86	1	12.20	22.48	1.30	1000	2158	1553	1401

Table A.5: Statistics for the 10-hour fine-tuning subsets selected in a purely random way

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
SPK_DIV_RND_24_0	48.55	9.16	8	16	24	97507	10385	2879	33.87	64	3	12.52	17.19	1.99	35	51	1117	1762
SPK_DIV_RND_24_1	50.46	9.13	11	13	24	96505	10814	2870	33.63	66	2	12.55	18.24	1.83	35	58	1825	1045
SPK_DIV_RND_24_3	48.10	9.16	13	11	24	98980	10268	2949	33.56	68	1	12.21	19.64	1.35	40	66	900	2049
SPK_DIV_RND_24_4	49.99	9.14	15	9	24	94185	10060	3008	31.31	61	1	11.97	17.31	1.24	34	60	1309	1699
SPK_DIV_RND_24_5	48.96	9.16	10	14	24	97200	10699	2949	32.96	63	1	12.22	17.23	1.51	40	66	1267	1682
SPK_DIV_RND_24_6	48.04	9.16	10	14	24	95412	10317	2907	32.82	60	1	12.39	17.19	1.87	32	58	1259	1648
SPK_DIV_RND_24_7	47.61	9.15	17	7	24	94306	10498	2980	31.65	64	1	12.09	17.19	1.32	34	64	724	2256
SPK_DIV_RND_24_8	47.08	9.17	10	14	24	97316	9887	2836	34.31	64	2	12.70	17.11	1.24	40	65	981	1855
SPK_DIV_RND_96_0	50.40	9.14	39	57	96	99173	11619	2929	33.86	69	1	12.31	17.15	1.08	131	217	1426	1503
SPK_DIV_RND_96_1	49.52	9.15	49	47	96	96915	11382	2956	32.79	67	1	12.18	18.78	1.49	134	226	1488	1468
SPK_DIV_RND_96_2	51.61	9.12	49	47	96	97145	11304	3023	32.14	74	1	11.92	27.14	0.83	134	216	1781	1242
SPK_DIV_RND_96_3	51.26	9.13	44	52	96	96593	11322	2975	32.47	66	1	12.10	20.65	1.29	141	219	1689	1286
SPK_DIV_RND_96_4	51.18	9.13	47	49	96	98004	11413	2959	33.12	67	1	12.19	17.22	1.19	145	219	1624	1335
SPK_DIV_RND_96_5	48.72	9.16	43	53	96	97886	11325	2972	32.94	64	1	12.12	18.78	1.08	114	231	1302	1670
SPK_DIV_RND_96_6	50.63	9.15	46	50	96	96934	11226	3031	31.98	80	1	11.89	22.80	1.49	119	225	1294	1737
SPK_DIV_RND_96_7	49.48	9.16	45	51	96	96133	11471	2987	32.18	67	1	12.05	20.00	1.32	108	237	1116	1871

Table A.6: Statistics for the 10-hour fine-tuning subsets selected to test the impact of speaker diversity

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
UTTLN_DIV_RND_LNG_DUR_TAIL_0	49.53	9.15	699	751	1450	98120	12146	2254	43.53	70	21	16.00	21.39	15.54	904	1788	1151	1103
UTTLN_DIV_RND_LNG_DUR_TAIL_1	49.51	9.14	693	740	1433	98145	12197	2252	43.58	86	14	16.02	22.80	15.54	915	1774	1156	1096
UTTLN_DIV_RND_LNG_DUR_TAIL_2	49.35	9.14	682	725	1407	98098	12149	2252	43.56	80	17	16.01	29.74	15.54	910	1778	1156	1096
UTTLN_DIV_RND_LNG_DUR_TAIL_3	49.43	9.15	676	739	1415	97894	12345	2254	43.43	73	19	16.00	22.72	15.54	916	1763	1156	1098
UTTLN_DIV_RND_LNG_DUR_TAIL_4	49.36	9.15	701	732	1433	98622	12311	2253	43.77	75	17	16.01	21.63	15.54	902	1773	1111	1142
UTTLN_DIV_RND_LNG_DUR_TAIL_5	49.41	9.14	691	724	1415	98455	12331	2254	43.68	72	16	16.00	25.85	15.54	892	1758	1172	1082
UTTLN_DIV_RND_LNG_DUR_TAIL_6	49.25	9.15	686	719	1405	98304	12195	2255	43.59	74	9	16.00	23.39	15.54	890	1753	1101	1154
UTTLN_DIV_RND_LNG_DUR_TAIL_7	49.27	9.14	684	730	1414	97847	12215	2251	43.47	70	16	16.02	27.92	15.54	911	1770	1141	1110
UTTLN_DIV_RND_MIDDLE_DUR_0	48.54	9.16	752	806	1558	97884	12379	2619	37.37	59	12	13.77	14.23	13.24	950	1984	1299	1320
UTTLN_DIV_RND_MIDDLE_DUR_1	48.47	9.16	728	813	1541	97917	12181	2618	37.40	61	15	13.78	14.23	13.24	977	1977	1328	1290
UTTLN_DIV_RND_MIDDLE_DUR_2	48.77	9.16	742	813	1555	97633	12272	2619	37.28	61	11	13.77	14.23	13.24	961	2010	1312	1307
UTTLN_DIV_RND_MIDDLE_DUR_3	48.56	9.16	742	801	1543	97910	12390	2616	37.43	61	12	13.77	14.23	13.24	973	2000	1309	1307
UTTLN_DIV_RND_MIDDLE_DUR_4	48.46	9.16	744	796	1540	97696	12277	2617	37.33	58	13	13.77	14.23	13.24	970	1986	1328	1289
UTTLN_DIV_RND_MIDDLE_DUR_5	49.04	9.16	746	809	1555	97779	12430	2616	37.38	60	12	13.77	14.23	13.24	979	2021	1317	1299
UTTLN_DIV_RND_MIDDLE_DUR_6	48.61	9.16	742	811	1553	97376	12292	2616	37.22	62	11	13.77	14.23	13.24	971	2012	1318	1298
UTTLN_DIV_RND_MIDDLE_DUR_7	48.61	9.16	720	826	1546	97588	12563	2617	37.29	63	10	13.77	14.23	13.24	969	2023	1311	1306
UTTLN_DIV_RND_SHRT_DUR_TAIL_0	56.30	9.10	959	1054	2013	98214	12774	7767	12.65	32	1	4.64	7.32	0.83	1187	3457	4729	3038
UTTLN_DIV_RND_SHRT_DUR_TAIL_1	56.07	9.10	980	1046	2026	98126	12575	7763	12.64	31	1	4.65	7.32	0.92	1183	3417	4688	3075
UTTLN_DIV_RND_SHRT_DUR_TAIL_2	56.99	9.10	957	1060	2017	98203	12583	7756	12.66	33	1	4.65	7.32	1.13	1179	3421	4713	3043
UTTLN_DIV_RND_SHRT_DUR_TAIL_3	56.57	9.10	943	1055	1998	98182	12687	7797	12.59	32	1	4.63	7.32	0.83	1174	3383	4732	3065
UTTLN_DIV_RND_SHRT_DUR_TAIL_4	56.63	9.10	974	1035	2009	98258	12814	7725	12.72	32	1	4.67	7.32	1.11	1181	3421	4660	3065
UTTLN_DIV_RND_SHRT_DUR_TAIL_5	56.60	9.10	976	1048	2024	98384	12567	7724	12.74	32	1	4.67	7.32	0.83	1184	3414	4679	3045
UTTLN_DIV_RND_SHRT_DUR_TAIL_6	56.50	9.10	965	1062	2027	98062	12592	7823	12.54	32	1	4.61	7.32	0.92	1195	3465	4735	3088
UTTLN_DIV_RND_SHRT_DUR_TAIL_7	56.45	9.10	950	1054	2004	98534	12633	7747	12.72	31	1	4.66	7.32	1.16	1167	3408	4689	3058

Table A.7: Statistics for the 10-hour fine-tuning subsets selected to test the impact of utterance length

A.2 Statistics for the 1-hour Subsets

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
PERPLEXITY_5k_LM_15_HEAD_0	35.50	9.19	138	132	270	9452	2408	294	32.15	57	4	12.27	16.79	1.97	238	281	101	193
PERPLEXITY_5k_LM_15_HEAD_1	35.18	9.19	132	135	267	9605	2469	294	32.67	56	5	12.63	16.75	2.06	238	281	99	195
PERPLEXITY_5k_LM_15_HEAD_2	35.07	9.20	138	126	264	9601	2569	297	32.33	55	3	12.35	16.93	2.01	232	282	101	196
PERPLEXITY_5k_LM_15_HEAD_3	35.26	9.19	146	125	271	9542	2501	293	32.57	58	3	12.37	17.02	2.14	228	279	92	201
PERPLEXITY_5k_LM_15_HEAD_4	35.25	9.19	143	128	271	9188	2449	293	31.36	56	5	12.08	16.58	2.33	226	278	106	187
PERPLEXITY_5k_LM_15_HEAD_5	35.26	9.19	145	129	274	9422	2579	294	32.05	56	3	12.33	17.23	1.91	229	284	91	203
PERPLEXITY_5k_LM_15_HEAD_6	35.26	9.20	121	136	257	9272	2532	295	31.43	60	4	12.31	16.62	1.81	223	277	104	191
PERPLEXITY_5k_LM_15_HEAD_7	35.25	9.19	132	137	269	9332	2466	297	31.42	63	3	12.03	16.80	1.64	233	280	113	184
PERPLEXITY_5k_LM_15_TAIL_0	72.05	9.07	129	159	288	9761	3138	348	28.05	62	3	10.28	16.80	1.77	265	319	256	92
PERPLEXITY_5k_LM_15_TAIL_1	71.56	9.07	150	160	310	10204	3104	348	29.32	61	3	10.45	17.05	1.77	267	326	245	103
PERPLEXITY_5k_LM_15_TAIL_2	72.87	9.06	130	157	287	10026	3123	346	28.98	64	2	10.34	17.05	1.46	261	319	260	86
PERPLEXITY_5k_LM_15_TAIL_3	71.98	9.07	147	152	299	10386	3145	346	30.02	60	2	10.70	16.97	1.96	270	324	264	82
PERPLEXITY_5k_LM_15_TAIL_4	73.64	9.05	153	137	290	10312	3186	345	29.89	63	1	10.59	16.78	1.39	265	321	241	104
PERPLEXITY_5k_LM_15_TAIL_5	72.75	9.07	129	160	289	9980	3075	341	29.27	67	3	10.40	16.90	1.94	266	318	264	77
PERPLEXITY_5k_LM_15_TAIL_6	71.72	9.08	151	154	305	9912	3150	345	28.73	75	2	10.47	24.53	1.71	275	329	245	100
PERPLEXITY_5k_LM_15_TAIL_7	73.97	9.06	137	150	287	9939	3035	345	28.81	65	3	10.30	17.00	1.90	246	316	264	81

Table A.8: Statistics for the 1-hour fine-tuning subsets selected based on perplexity of Byte Pair Encoded (BPE) clustered units

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
PRETRAIN_M_LOSS_AVG_ASC_0	64.87	8.98	205	221	426	10297	3089	562	18.32	62	1	6.48	19.87	1.51	357	478	423	139
PRETRAIN_M_LOSS_AVG_ASC_1	64.74	8.98	222	206	428	9994	3071	569	17.56	64	1	6.23	16.74	1.04	364	479	433	136
PRETRAIN_M_LOSS_AVG_ASC_2	64.50	8.98	224	218	442	10463	3056	571	18.32	64	1	6.39	16.52	1.11	378	497	440	131
PRETRAIN_M_LOSS_AVG_ASC_3	65.29	8.98	208	221	429	10289	3068	570	18.05	58	1	6.33	16.98	1.19	350	484	437	133
PRETRAIN_M_LOSS_AVG_ASC_4	64.37	8.98	208	208	416	10244	3096	564	18.16	58	1	6.45	16.89	1.15	350	477	435	129
PRETRAIN_M_LOSS_AVG_ASC_5	64.81	8.97	207	211	418	10020	3080	568	17.64	55	1	6.28	16.98	1.25	343	474	434	134
PRETRAIN_M_LOSS_AVG_ASC_6	65.60	8.98	214	209	423	10467	3174	561	18.66	56	1	6.51	17.18	1.07	357	469	432	129
PRETRAIN_M_LOSS_AVG_ASC_7	64.99	8.97	196	204	400	10176	3058	564	18.04	59	1	6.38	16.65	1.17	351	467	447	117
PRETRAIN_M_LOSS_AVG_DESC_0	43.28	9.29	135	150	285	8354	2213	352	23.73	52	1	10.17	16.45	1.07	241	307	121	231
PRETRAIN_M_LOSS_AVG_DESC_1	42.47	9.29	113	158	271	9016	2462	355	25.40	50	1	10.78	17.06	1.33	217	300	107	248
PRETRAIN_M_LOSS_AVG_DESC_2	41.95	9.30	136	154	290	8422	2295	349	24.13	51	3	10.22	16.89	1.57	233	310	111	238
PRETRAIN_M_LOSS_AVG_DESC_3	44.04	9.29	131	144	275	8237	2378	356	23.14	58	2	10.14	16.20	1.79	243	303	115	241
PRETRAIN_M_LOSS_AVG_DESC_4	41.82	9.29	132	150	282	8359	2339	354	23.61	53	1	10.20	16.87	1.26	227	311	109	245
PRETRAIN_M_LOSS_AVG_DESC_5	42.28	9.30	127	163	290	8492	2357	353	24.06	50	1	10.34	16.77	1.13	236	305	115	238
PRETRAIN_M_LOSS_AVG_DESC_6	42.42	9.29	135	138	273	8212	2251	348	23.60	48	1	10.17	16.93	1.32	246	305	119	229
PRETRAIN_M_LOSS_AVG_DESC_7	43.24	9.30	127	156	283	8502	2282	350	24.29	50	1	10.24	16.96	1.43	246	310	118	232
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_0	56.93	8.97	156	130	286	9553	2679	336	28.43	62	2	10.67	17.07	1.62	251	308	263	73
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_1	59.04	8.96	156	130	286	9560	2785	340	28.12	56	1	10.54	17.01	1.48	258	311	278	62
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_2	58.67	8.96	149	137	286	9558	2768	335	28.53	63	1	10.71	17.11	1.24	250	312	272	63
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_3	58.06	8.95	152	142	294	9507	2717	336	28.29	58	2	10.68	17.13	1.91	252	315	278	58
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_4	59.87	8.95	152	131	283	10046	2837	335	29.99	59	4	11.20	16.74	1.98	243	299	281	54
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_5	58.87	8.96	149	132	281	9857	2775	332	29.69	57	2	11.13	17.08	1.70	253	305	273	59
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_6	59.28	8.96	171	125	296	9452	2752	336	28.13	59	2	10.57	16.75	1.61	254	314	276	60
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_7	58.01	8.95	156	132	288	9672	2836	338	28.62	57	1	10.74	16.89	1.45	249	317	282	56
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_0	44.84	9.31	115	145	260	9830	2755	291	33.78	61	3	12.39	17.20	2.04	232	273	82	209
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_1	45.90	9.31	113	146	259	9604	2701	294	32.67	59	5	11.95	16.77	2.07	227	272	85	209
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_2	44.67	9.30	118	143	261	9750	2795	293	33.28	63	2	12.21	16.84	2.25	234	275	83	210
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_3	45.02	9.31	97	166	263	9779	2730	296	33.04	66	6	12.11	16.70	2.53	243	273	75	221
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_4	44.78	9.30	127	146	273	9586	2783	295	32.49	54	2	11.97	17.01	2.18	230	282	94	201
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_5	44.82	9.31	113	153	266	9491	2759	294	32.28	59	4	12.26	16.86	1.89	236	279	89	205
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_6	45.16	9.30	115	144	259	9756	2769	293	33.30	61	3	12.31	17.08	1.58	231	275	85	208
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_7	45.36	9.30	119	149	268	9919	2857	293	33.85	61	5	12.47	16.87	2.06	236	280	71	222

Table A.9: Statistics for the 1-hour fine-tuning subsets selected based on pre-training loss

Subset	Average perplexity	Average unmasked loss	Number of female speakers	Number of male speakers	Total speakers	Total tokens	Unique vocabulary words	Total utterances	Average utterance length	Maximum utterance length	Minimum utterance length	Average utterance duration	Maximum utterance duration	Minimum utterance duration	Number of books	Number of chapters	utterances from train-other	utterances from train-clean
PUR_RND_0	48.83	9.14	128	144	272	9558	2760	292	32.73	58	1	12.10	16.63	1.62	233	284	164	128
PUR_RND_1	48.38	9.15	123	144	267	10134	2859	291	34.82	62	4	12.67	16.96	1.99	241	281	154	137
PUR_RND_2	50.89	9.15	136	138	274	9542	2770	292	32.68	59	4	11.92	16.93	1.97	244	284	156	136
PUR_RND_3	49.70	9.14	124	150	274	9854	2833	292	33.75	59	4	12.42	16.56	1.94	246	283	145	147
PUR_RND_4	49.69	9.14	132	146	278	9665	2818	292	33.10	67	3	12.32	17.32	1.78	246	283	146	146
PUR_RND_5	50.23	9.13	132	136	268	10123	2854	292	34.67	60	4	12.53	16.86	2.00	227	281	147	145
PUR_RND_6	50.76	9.14	129	144	273	10008	2955	292	34.27	57	4	12.61	17.01	2.24	242	282	147	145
PUR_RND_7	50.48	9.14	133	143	276	10103	2912	292	34.60	62	5	12.56	17.31	2.27	242	285	158	134

Table A.10: Statistics for the 1-hour fine-tuning subsets selected in a purely random way

Appendix B: Results Breakdown for the Different Data Selection Criteria

This appendix provides a detailed breakdown of the WER obtained on Librispeech test-other and Librispeech test-clean when fine-tuning using each data selection criterion. The result for each of the 8 runs belonging to a particular criterion is presented, then the minimum, maximum, mean and standard deviation of the WER are demonstrated to summarize the results of all the data selection criteria.

B.1 Results Breakdown on Librispeech Test-other

This section breaks down the WER results on test-other achieved when fine-tuning the HuBERT base model using each selection criterion. It demonstrates the results of pure random data selection. Moreover, it shows the impact of book diversity, gender bias, speaker diversity, utterance duration as well the proposed data selection criterion in this work: PBPE and pre-training loss on the downstream WER.

B.1.1 Results Obtained when Fine-tuning with 10-hour Subsets

run	WER
PUR_RND_10	8.83
PUR_RND_11	9.12
PUR_RND_13	9.04
PUR_RND_14	9.04
PUR_RND_4	9.03
PUR_RND_6	8.95
PUR_RND_7	8.95
PUR_RND_9	9.03

Table B.1: WER on Librispeech test-other for pure-random data selection in the 10-hour setup

run	WER
BK_DIV_RND_16_0	9.69
BK_DIV_RND_16_1	9.52
BK_DIV_RND_16_2	9.48
BK_DIV_RND_16_3	9.67
BK_DIV_RND_16_4	9.28
BK_DIV_RND_16_5	9.55
BK_DIV_RND_16_6	9.57
BK_DIV_RND_16_7	9.37
BK_DIV_RND_64_0	9.28
BK_DIV_RND_64_1	9.11
BK_DIV_RND_64_2	9.03
BK_DIV_RND_64_3	9.23
BK_DIV_RND_64_4	9.36
BK_DIV_RND_64_5	9.36
BK_DIV_RND_64_6	9.16
BK_DIV_RND_64_7	9.09

Table B.2: WER on Librispeech test-other when fixing the number of audio-books during data selection in the 10-hour setup

run	WER
GNDR_DIV_F_24_0	9.68
GNDR_DIV_F_24_2	10.14
GNDR_DIV_F_24_4	9.5
GNDR_DIV_F_24_5	9.72
GNDR_DIV_F_24_6	9.56
GNDR_DIV_F_24_7	9.61
GNDR_DIV_F_24_8	9.77
GNDR_DIV_F_24_9	9.57
GNDR_DIV_M_24_0	9.63
GNDR_DIV_M_24_1	9.44
GNDR_DIV_M_24_2	9.61
GNDR_DIV_M_24_3	9.44
GNDR_DIV_M_24_4	9.81
GNDR_DIV_M_24_5	9.75
GNDR_DIV_M_24_6	9.58
GNDR_DIV_M_24_7	9.78

Table B.3: WER on Librispeech test-other when biasing the selected subset to a particular gender in the 10-hour setup

run	WER
SPK_DIV_RND_24_0	9.99
SPK_DIV_RND_24_1	9.58
SPK_DIV_RND_24_3	9.75
SPK_DIV_RND_24_4	9.8
SPK_DIV_RND_24_5	9.66
SPK_DIV_RND_24_6	9.44
SPK_DIV_RND_24_7	9.92
SPK_DIV_RND_24_8	9.52
SPK_DIV_RND_96_0	9.09
SPK_DIV_RND_96_1	9.35
SPK_DIV_RND_96_2	9.01
SPK_DIV_RND_96_3	9.04
SPK_DIV_RND_96_4	9.07
SPK_DIV_RND_96_5	9.24
SPK_DIV_RND_96_6	9.1
SPK_DIV_RND_96_7	9.24

Table B.4: WER on Librispeech test-other when fixing the number of speakers during data selection in the 10-hour setup

run	WER
PERPLEXITY_5k_LM_15_HEAD_0	9.33
PERPLEXITY_5k_LM_15_HEAD_10	9.45
PERPLEXITY_5k_LM_15_HEAD_13	9.39
PERPLEXITY_5k_LM_15_HEAD_14	9.23
PERPLEXITY_5k_LM_15_HEAD_3	9.2
PERPLEXITY_5k_LM_15_HEAD_4	9.46
PERPLEXITY_5k_LM_15_HEAD_7	9.4
PERPLEXITY_5k_LM_15_HEAD_9	9.43
PERPLEXITY_5k_LM_15_TAIL_0	8.81
PERPLEXITY_5k_LM_15_TAIL_2	9
PERPLEXITY_5k_LM_15_TAIL_3	8.85
PERPLEXITY_5k_LM_15_TAIL_4	8.88
PERPLEXITY_5k_LM_15_TAIL_5	8.93
PERPLEXITY_5k_LM_15_TAIL_7	9.04
PERPLEXITY_5k_LM_15_TAIL_8	8.94
PERPLEXITY_5k_LM_15_TAIL_9	8.97
PERPLEXITY_5k_LM_40_MIDDLE_0	9.07
PERPLEXITY_5k_LM_40_MIDDLE_1	8.85
PERPLEXITY_5k_LM_40_MIDDLE_2	9.08
PERPLEXITY_5k_LM_40_MIDDLE_3	8.97
PERPLEXITY_5k_LM_40_MIDDLE_4	8.85
PERPLEXITY_5k_LM_40_MIDDLE_5	9
PERPLEXITY_5k_LM_40_MIDDLE_6	9.14
PERPLEXITY_5k_LM_40_MIDDLE_7	9.12

Table B.5: WER on Librispeech test-other for data selection based on the perplexity of byte pair encoded clustered units (PBPE) in the 10-hour setup

run	WER
PRETRAIN_M_LOSS_AVG_ASC_0	8.93
PRETRAIN_M_LOSS_AVG_ASC_1	9.08
PRETRAIN_M_LOSS_AVG_ASC_2	9.05
PRETRAIN_M_LOSS_AVG_ASC_3	9.1
PRETRAIN_M_LOSS_AVG_ASC_4	9.08
PRETRAIN_M_LOSS_AVG_ASC_5	8.98
PRETRAIN_M_LOSS_AVG_ASC_6	8.92
PRETRAIN_M_LOSS_AVG_ASC_7	8.88
PRETRAIN_M_LOSS_AVG_DESC_0	9.28
PRETRAIN_M_LOSS_AVG_DESC_1	9.17
PRETRAIN_M_LOSS_AVG_DESC_2	9.42
PRETRAIN_M_LOSS_AVG_DESC_3	9.49
PRETRAIN_M_LOSS_AVG_DESC_4	9.26
PRETRAIN_M_LOSS_AVG_DESC_5	9.5
PRETRAIN_M_LOSS_AVG_DESC_6	9.42
PRETRAIN_M_LOSS_AVG_DESC_7	9.33
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_0	8.95
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_1	9.04
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_2	9.05
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_3	9.32
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_4	9.1
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_5	9.2
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_6	9.14
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_7	9.01
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_0	9.15
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_1	9.16
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_2	9.06
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_3	9.09
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_4	9.15
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_5	9.28
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_6	9.26
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_7	9.14

Table B.6: WER on Librispeech test-other for data selection based on pre-training loss in the 10-hour setup

run	WER
UTTLN_DIV_RND_LNG_DUR_TAIL_0	9.07
UTTLN_DIV_RND_LNG_DUR_TAIL_1	9.14
UTTLN_DIV_RND_LNG_DUR_TAIL_2	9.09
UTTLN_DIV_RND_LNG_DUR_TAIL_3	9.18
UTTLN_DIV_RND_LNG_DUR_TAIL_4	9.08
UTTLN_DIV_RND_LNG_DUR_TAIL_5	8.94
UTTLN_DIV_RND_LNG_DUR_TAIL_6	8.95
UTTLN_DIV_RND_LNG_DUR_TAIL_7	9.18
UTTLN_DIV_RND_MIDDLE_DUR_0	9.2
UTTLN_DIV_RND_MIDDLE_DUR_1	9.03
UTTLN_DIV_RND_MIDDLE_DUR_2	9.02
UTTLN_DIV_RND_MIDDLE_DUR_3	9.36
UTTLN_DIV_RND_MIDDLE_DUR_4	9.09
UTTLN_DIV_RND_MIDDLE_DUR_5	9.24
UTTLN_DIV_RND_MIDDLE_DUR_6	8.96
UTTLN_DIV_RND_MIDDLE_DUR_7	9.28
UTTLN_DIV_RND_SHRT_DUR_TAIL_0	8.99
UTTLN_DIV_RND_SHRT_DUR_TAIL_1	9.02
UTTLN_DIV_RND_SHRT_DUR_TAIL_2	8.97
UTTLN_DIV_RND_SHRT_DUR_TAIL_3	8.95
UTTLN_DIV_RND_SHRT_DUR_TAIL_4	8.99
UTTLN_DIV_RND_SHRT_DUR_TAIL_5	9.04
UTTLN_DIV_RND_SHRT_DUR_TAIL_6	8.83
UTTLN_DIV_RND_SHRT_DUR_TAIL_7	9.09

Table B.7: WER on Librispeech test-other based on utterance duration in the 10-hour setup

Criterion	Mean WER	StdDev of WER	Min WER	Max WER
PERPLEXITY_5k_LM_15_TAIL	8.93	0.08	8.81	9.04
UTTLN_DIV_RND_SHRT_DUR_TAIL	8.99	0.08	8.83	9.09
PUR_RND	9.00	0.09	8.83	9.12
PRETRAIN_M_LOSS_AVG_ASC	9.00	0.09	8.88	9.1
PERPLEXITY_5k_LM_40_MIDDLE	9.01	0.11	8.85	9.14
UTTLN_DIV_RND_LNG_DUR_TAIL	9.08	0.09	8.94	9.18
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD	9.10	0.12	8.95	9.32
SPK_DIV_RND_96	9.14	0.12	9.01	9.35
UTTLN_DIV_RND_MIDDLE_DUR	9.15	0.14	8.96	9.36
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL	9.16	0.08	9.06	9.28
BK_DIV_RND_64	9.20	0.13	9.03	9.36
PRETRAIN_M_LOSS_AVG_DESC	9.36	0.12	9.17	9.5
PERPLEXITY_5k_LM_15_HEAD	9.36	0.1	9.2	9.46
BK_DIV_RND_16	9.52	0.14	9.28	9.69
GNDR_DIV_M_24	9.63	0.14	9.44	9.81
GNDR_DIV_F_24	9.69	0.20	9.5	10.14
SPK_DIV_RND_24	9.71	0.19	9.44	9.99

Table B.8: Librispeech test-other WER Summary for different data selection criteria in the 10-hour setup

B.1.2 Results Obtained when Fine-tuning with 1-hour Subsets

run	WER
PUR_RND_0	12
PUR_RND_1	12
PUR_RND_2	12.05
PUR_RND_3	12.13
PUR_RND_4	11.88
PUR_RND_5	12.19
PUR_RND_6	11.85
PUR_RND_7	11.77

Table B.9: WER on Librispeech test-other for pure-random data selection in the 1-hour setup

run	WER
PERPLEXITY_5k_LM_15_HEAD_0	12.34
PERPLEXITY_5k_LM_15_HEAD_1	12.27
PERPLEXITY_5k_LM_15_HEAD_2	11.9
PERPLEXITY_5k_LM_15_HEAD_3	12.43
PERPLEXITY_5k_LM_15_HEAD_4	12.29
PERPLEXITY_5k_LM_15_HEAD_5	12.21
PERPLEXITY_5k_LM_15_HEAD_6	12.32
PERPLEXITY_5k_LM_15_HEAD_7	12.6
PERPLEXITY_5k_LM_15_TAIL_0	12.12
PERPLEXITY_5k_LM_15_TAIL_1	11.78
PERPLEXITY_5k_LM_15_TAIL_2	12.03
PERPLEXITY_5k_LM_15_TAIL_3	12.04
PERPLEXITY_5k_LM_15_TAIL_4	12.11
PERPLEXITY_5k_LM_15_TAIL_5	12.39
PERPLEXITY_5k_LM_15_TAIL_6	12.01
PERPLEXITY_5k_LM_15_TAIL_7	12.2

Table B.10: WER on Librispeech test-other for data selection based on the perplexity of byte pair encoded clustered units (PBPE) in the 1-hour setup

run	WER
PRETRAIN_M_LOSS_AVG_ASC_0	12.06
PRETRAIN_M_LOSS_AVG_ASC_1	12.07
PRETRAIN_M_LOSS_AVG_ASC_2	12.19
PRETRAIN_M_LOSS_AVG_ASC_3	12
PRETRAIN_M_LOSS_AVG_ASC_4	12.11
PRETRAIN_M_LOSS_AVG_ASC_5	11.85
PRETRAIN_M_LOSS_AVG_ASC_6	11.97
PRETRAIN_M_LOSS_AVG_ASC_7	12.04
PRETRAIN_M_LOSS_AVG_DESC_0	12.55
PRETRAIN_M_LOSS_AVG_DESC_1	12.12
PRETRAIN_M_LOSS_AVG_DESC_2	12.67
PRETRAIN_M_LOSS_AVG_DESC_3	12.43
PRETRAIN_M_LOSS_AVG_DESC_4	12.26
PRETRAIN_M_LOSS_AVG_DESC_5	12.49
PRETRAIN_M_LOSS_AVG_DESC_6	12.55
PRETRAIN_M_LOSS_AVG_DESC_7	12.54
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_0	12.6
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_1	12.28
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_2	11.97
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_3	12.22
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_4	12.15
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_5	12.51
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_6	12.19
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_7	12.22
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_0	12.35
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_1	12.29
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_2	12.77
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_3	12.18
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_4	12.31
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_5	12.1
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_6	12.15
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_7	12.25

Table B.11: WER on Librispeech test-other for data selection based on pre-training loss in the 1-hour setup

Criterion	Mean WER	StdDev of WER	Min WER	Max WER
PUR_RND	11.98	0.14	11.77	12.19
PRETRAIN_M_LOSS_AVG_ASC	12.04	0.10	11.85	12.19
PERPLEXITY_5k_LM_15_TAIL	12.09	0.17	11.78	12.39
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD	12.27	0.20	11.97	12.6
PERPLEXITY_5k_LM_15_HEAD	12.30	0.20	11.9	12.6
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL	12.30	0.21	12.1	12.77
PRETRAIN_M_LOSS_AVG_DESC	12.45	0.18	12.12	12.67

Table B.12: Librispeech test-other WER Summary for different data selection criteria in the 1-hour setup

B.2 Results Breakdown on Librispeech Test-clean

This section breaks down the WER results on test-clean achieved when fine-tuning the HuBERT base model using each selection criterion. Similar to the previous section, it demonstrates the results of pure random data selection. Moreover, it shows the impact of book diversity, gender bias, speaker diversity, utterance duration as well as the proposed data selection criterion in this work: PBPE and pre-training loss on the downstream WER.

B.2.1 Results Obtained when Fine-tuning with 10-hour Subsets

run	WER
PUR_RND_10	4.33
PUR_RND_11	4.2
PUR_RND_13	4.25
PUR_RND_14	4.21
PUR_RND_4	4.22
PUR_RND_6	4.26
PUR_RND_7	4.18
PUR_RND_9	4.26

Table B.13: WER on Librispeech test-clean for pure-random data selection in the 10-hour setup

run	WER
BK_DIV_RND_16_0	4.58
BK_DIV_RND_16_1	4.62
BK_DIV_RND_16_2	4.44
BK_DIV_RND_16_3	4.69
BK_DIV_RND_16_4	4.54
BK_DIV_RND_16_5	4.56
BK_DIV_RND_16_6	4.69
BK_DIV_RND_16_7	4.45
BK_DIV_RND_64_0	4.36
BK_DIV_RND_64_1	4.38
BK_DIV_RND_64_2	4.33
BK_DIV_RND_64_3	4.41
BK_DIV_RND_64_4	4.41
BK_DIV_RND_64_5	4.45
BK_DIV_RND_64_6	4.4
BK_DIV_RND_64_7	4.31

Table B.14: WER on Librispeech test-clean when fixing the number of audiobooks during data selection in the 10-hour setup

run	WER
GNDR_DIV_F_24_0	4.62
GNDR_DIV_F_24_2	4.6
GNDR_DIV_F_24_4	4.53
GNDR_DIV_F_24_5	4.48
GNDR_DIV_F_24_6	4.5
GNDR_DIV_F_24_7	4.56
GNDR_DIV_F_24_8	4.47
GNDR_DIV_F_24_9	4.56
GNDR_DIV_M_24_0	4.45
GNDR_DIV_M_24_1	4.53
GNDR_DIV_M_24_2	4.56
GNDR_DIV_M_24_3	4.53
GNDR_DIV_M_24_4	4.63
GNDR_DIV_M_24_5	4.47
GNDR_DIV_M_24_6	4.69
GNDR_DIV_M_24_7	4.48

Table B.15: WER on Librispeech test-clean when biasing the selected subset to a particular gender in the 10-hour setup

run	WER
SPK_DIV_RND_24_0	4.35
SPK_DIV_RND_24_1	4.53
SPK_DIV_RND_24_3	4.48
SPK_DIV_RND_24_4	4.53
SPK_DIV_RND_24_5	4.55
SPK_DIV_RND_24_6	4.58
SPK_DIV_RND_24_7	4.69
SPK_DIV_RND_24_8	4.49
SPK_DIV_RND_96_0	4.24
SPK_DIV_RND_96_1	4.51
SPK_DIV_RND_96_2	4.28
SPK_DIV_RND_96_3	4.42
SPK_DIV_RND_96_4	4.3
SPK_DIV_RND_96_5	4.34
SPK_DIV_RND_96_6	4.38
SPK_DIV_RND_96_7	4.33

Table B.16: WER on Librispeech test-clean when fixing the number of speakers during data selection in the 10-hour setup

run	WER
PERPLEXITY_5k_LM_15_HEAD_0	4.38
PERPLEXITY_5k_LM_15_HEAD_10	4.37
PERPLEXITY_5k_LM_15_HEAD_13	4.55
PERPLEXITY_5k_LM_15_HEAD_14	4.36
PERPLEXITY_5k_LM_15_HEAD_3	4.5
PERPLEXITY_5k_LM_15_HEAD_4	4.36
PERPLEXITY_5k_LM_15_HEAD_7	4.35
PERPLEXITY_5k_LM_15_HEAD_9	4.51
PERPLEXITY_5k_LM_15_TAIL_0	4.27
PERPLEXITY_5k_LM_15_TAIL_2	4.26
PERPLEXITY_5k_LM_15_TAIL_3	4.21
PERPLEXITY_5k_LM_15_TAIL_4	4.27
PERPLEXITY_5k_LM_15_TAIL_5	4.22
PERPLEXITY_5k_LM_15_TAIL_7	4.24
PERPLEXITY_5k_LM_15_TAIL_8	4.29
PERPLEXITY_5k_LM_15_TAIL_9	4.27
PERPLEXITY_5k_LM_40_MIDDLE_0	4.25
PERPLEXITY_5k_LM_40_MIDDLE_1	4.17
PERPLEXITY_5k_LM_40_MIDDLE_2	4.19
PERPLEXITY_5k_LM_40_MIDDLE_3	4.23
PERPLEXITY_5k_LM_40_MIDDLE_4	4.21
PERPLEXITY_5k_LM_40_MIDDLE_5	4.25
PERPLEXITY_5k_LM_40_MIDDLE_6	4.33
PERPLEXITY_5k_LM_40_MIDDLE_7	4.27

Table B.17: WER on Librispeech test-clean for data selection based on the perplexity of byte pair encoded clustered units (PBPE) in the 10-hour setup

run	WER
PRETRAIN_M_LOSS_AVG_ASC_0	4.25
PRETRAIN_M_LOSS_AVG_ASC_1	4.26
PRETRAIN_M_LOSS_AVG_ASC_2	4.32
PRETRAIN_M_LOSS_AVG_ASC_3	4.26
PRETRAIN_M_LOSS_AVG_ASC_4	4.29
PRETRAIN_M_LOSS_AVG_ASC_5	4.28
PRETRAIN_M_LOSS_AVG_ASC_6	4.3
PRETRAIN_M_LOSS_AVG_ASC_7	4.3
PRETRAIN_M_LOSS_AVG_DESC_0	4.39
PRETRAIN_M_LOSS_AVG_DESC_1	4.53
PRETRAIN_M_LOSS_AVG_DESC_2	4.53
PRETRAIN_M_LOSS_AVG_DESC_3	4.43
PRETRAIN_M_LOSS_AVG_DESC_4	4.43
PRETRAIN_M_LOSS_AVG_DESC_5	4.38
PRETRAIN_M_LOSS_AVG_DESC_6	4.52
PRETRAIN_M_LOSS_AVG_DESC_7	4.45
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_0	4.39
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_1	4.48
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_2	4.33
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_3	4.48
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_4	4.36
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_5	4.4
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_6	4.43
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_7	4.37
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_0	4.27
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_1	4.18
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_2	4.2
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_3	4.24
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_4	4.21
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_5	4.26
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_6	4.32
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_7	4.17

Table B.18: WER on Librispeech test-clean for data selection based on pre-training loss in the 10-hour setup

run	WER
UTTLN_DIV_RND_LNG_DUR_TAIL_0	4.31
UTTLN_DIV_RND_LNG_DUR_TAIL_1	4.37
UTTLN_DIV_RND_LNG_DUR_TAIL_2	4.35
UTTLN_DIV_RND_LNG_DUR_TAIL_3	4.33
UTTLN_DIV_RND_LNG_DUR_TAIL_4	4.23
UTTLN_DIV_RND_LNG_DUR_TAIL_5	4.17
UTTLN_DIV_RND_LNG_DUR_TAIL_6	4.26
UTTLN_DIV_RND_LNG_DUR_TAIL_7	4.25
UTTLN_DIV_RND_MIDDLE_DUR_0	4.37
UTTLN_DIV_RND_MIDDLE_DUR_1	4.2
UTTLN_DIV_RND_MIDDLE_DUR_2	4.19
UTTLN_DIV_RND_MIDDLE_DUR_3	4.24
UTTLN_DIV_RND_MIDDLE_DUR_4	4.27
UTTLN_DIV_RND_MIDDLE_DUR_5	4.25
UTTLN_DIV_RND_MIDDLE_DUR_6	4.3
UTTLN_DIV_RND_MIDDLE_DUR_7	4.35
UTTLN_DIV_RND_SHRT_DUR_TAIL_0	4.47
UTTLN_DIV_RND_SHRT_DUR_TAIL_1	4.49
UTTLN_DIV_RND_SHRT_DUR_TAIL_2	4.43
UTTLN_DIV_RND_SHRT_DUR_TAIL_3	4.43
UTTLN_DIV_RND_SHRT_DUR_TAIL_4	4.44
UTTLN_DIV_RND_SHRT_DUR_TAIL_5	4.41
UTTLN_DIV_RND_SHRT_DUR_TAIL_6	4.44
UTTLN_DIV_RND_SHRT_DUR_TAIL_7	4.44

Table B.19: WER on Librispeech test-clean based on utterance duration in the 10-hour setup

Criterion	Mean WER	StdDev of WER	Min WER	Max WER
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL	4.23	0.05	4.17	4.32
PERPLEXITY_5k_LM_40_MIDDLE	4.24	0.05	4.17	4.33
PUR_RND	4.24	0.05	4.18	4.33
PERPLEXITY_5k_LM_15_TAIL	4.25	0.03	4.21	4.29
UTTLN_DIV_RND_MIDDLE_DUR	4.27	0.07	4.19	4.37
PRETRAIN_M_LOSS_AVG_ASC	4.28	0.02	4.25	4.32
UTTLN_DIV_RND_LNG_DUR_TAIL	4.28	0.07	4.17	4.37
SPK_DIV_RND_96	4.35	0.09	4.24	4.51
BK_DIV_RND_64	4.38	0.05	4.31	4.45
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD	4.41	0.05	4.33	4.48
PERPLEXITY_5k_LM_15_HEAD	4.42	0.08	4.35	4.55
UTTLN_DIV_RND_SHRT_DUR_TAIL	4.44	0.03	4.41	4.49
PRETRAIN_M_LOSS_AVG_DESC	4.46	0.06	4.38	4.53
SPK_DIV_RND_24	4.53	0.10	4.35	4.69
GNDR_DIV_F_24	4.54	0.05	4.47	4.62
GNDR_DIV_M_24	4.54	0.08	4.45	4.69
BK_DIV_RND_16	4.57	0.10	4.44	4.69

Table B.20: Librispeech test-clean WER Summary for different data selection criteria in the 10-hour setup

B.2.2 Results Obtained when Fine-tuning with 1-hour Subsets

run	WER
PUR_RND_0	6.43
PUR_RND_1	6.42
PUR_RND_2	6.69
PUR_RND_3	6.76
PUR_RND_4	6.6
PUR_RND_5	6.69
PUR_RND_6	6.76
PUR_RND_7	6.55

Table B.21: WER on Librispeech test-clean for pure-random data selection in the 1-hour setup

run	WER
PERPLEXITY_5k_LM_15_HEAD_0	6.89
PERPLEXITY_5k_LM_15_HEAD_1	6.89
PERPLEXITY_5k_LM_15_HEAD_2	6.74
PERPLEXITY_5k_LM_15_HEAD_3	6.98
PERPLEXITY_5k_LM_15_HEAD_4	6.82
PERPLEXITY_5k_LM_15_HEAD_5	6.87
PERPLEXITY_5k_LM_15_HEAD_6	6.76
PERPLEXITY_5k_LM_15_HEAD_7	6.92
PERPLEXITY_5k_LM_15_TAIL_0	6.65
PERPLEXITY_5k_LM_15_TAIL_2	6.4
PERPLEXITY_5k_LM_15_TAIL_3	6.76
PERPLEXITY_5k_LM_15_TAIL_4	6.65
PERPLEXITY_5k_LM_15_TAIL_5	6.59
PERPLEXITY_5k_LM_15_TAIL_6	6.3
PERPLEXITY_5k_LM_15_TAIL_7	6.66

Table B.22: WER on Librispeech test-clean for data selection based on the perplexity of byte pair encoded clustered units (PBPE) in the 1-hour setup

run	WER
PRETRAIN_M_LOSS_AVG_ASC_0	6.65
PRETRAIN_M_LOSS_AVG_ASC_1	6.57
PRETRAIN_M_LOSS_AVG_ASC_2	6.8
PRETRAIN_M_LOSS_AVG_ASC_3	6.65
PRETRAIN_M_LOSS_AVG_ASC_4	6.69
PRETRAIN_M_LOSS_AVG_ASC_5	6.46
PRETRAIN_M_LOSS_AVG_ASC_6	6.65
PRETRAIN_M_LOSS_AVG_ASC_7	6.82
PRETRAIN_M_LOSS_AVG_DESC_0	7.1
PRETRAIN_M_LOSS_AVG_DESC_1	6.92
PRETRAIN_M_LOSS_AVG_DESC_2	7.09
PRETRAIN_M_LOSS_AVG_DESC_3	7.16
PRETRAIN_M_LOSS_AVG_DESC_4	7.01
PRETRAIN_M_LOSS_AVG_DESC_5	7.01
PRETRAIN_M_LOSS_AVG_DESC_6	7.25
PRETRAIN_M_LOSS_AVG_DESC_7	7.23
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_0	7.1
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_1	6.74
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_2	6.74
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_3	6.79
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_4	6.72
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_5	6.88
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_6	6.92
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD_7	6.77
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_0	6.52
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_1	6.84
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_2	6.99
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_3	6.74
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_4	6.7
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_5	6.55
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_6	6.66
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL_7	6.62

Table B.23: WER on Librispeech test-clean for data selection based on pre-training loss in the 1-hour setup

Criterion	Mean WER	StdDev of WER	Min WER	Max WER
PERPLEXITY_5k_LM_15_TAIL	6.57	0.16	6.3	6.76
PUR_RND	6.61	0.14	6.42	6.76
PRETRAIN_M_LOSS_AVG_ASC	6.66	0.12	6.46	6.82
PRETRAIN_U_LOSS_AVG_NO_MASK_TAIL	6.70	0.15	6.52	6.99
PRETRAIN_U_LOSS_AVG_NO_MASK_HEAD	6.83	0.13	6.72	7.1
PERPLEXITY_5k_LM_15_HEAD	6.86	0.08	6.74	6.98
PRETRAIN_M_LOSS_AVG_DESC	7.10	0.11	6.92	7.25

Table B.24: Librispeech test-clean WER Summary for different data selection criteria in the 1-hour setup

Bibliography

[1] Amber Afshan, Kshitiz Kumar, and Jian Wu. Sequence-level confidence classifier for asr utterance accuracy and application to acoustic models. In *INTERSPEECH 2021*, pages 4084–4088, 08 2021.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[3] Murali Karthick Baskar, Andrew Rosenberg, Bhuvana Ramabhadran, Yu Zhang, and Pedro Moreno. Ask2mask: Guided data selection for masked speech modeling. *arXiv preprint arXiv:2202.12719*, 2022.

[4] Jayadev Billa. Improving low-resource asr performance with untranscribed out-of-domain data. *arXiv preprint arXiv:2106.01227*, 2021.

[5] Meng Cai and Jia Liu. Maxout neurons for deep convolutional and lstm neural networks in speech recognition. *Speech Communication*, 77:53–64, 2016.

[6] Diamantino Caseiro and Isabel Trancoso. Large vocabulary continuous speech recognition using weighted finite-state transducers. In *Inter-*

national Conference for Natural Language Processing in Portugal, pages 91–99. Springer, 2002.

[7] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021.

[8] Hung-An Chang and James R. Glass. Hierarchical large-margin gaussian mixture models for phonetic classification. *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 272–277, 2007.

[9] Delphine Charlet. Confidence-measure-driven unsupervised incremental adaptation for hmm-based speech recognition. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 357–360. IEEE, 2001.

[10] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[11] Félix de Chaumont Quitry, Asa Oines, Pedro Moreno, and Eugene Weinstein. High quality agreement-based semi-supervised training data for

acoustic modeling. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 592–596, 2016.

[12] Mark Gales and Steve Young. *Application of Hidden Markov Models in Speech Recognition*. Now Foundations and Trends, 2008.

[13] Neeraj Gaur, Tongzhou Chen, Ehsan Variani, Parisa Haghani, Bhuvana Ramabhadran, and Pedro J. Moreno. Multilingual second-pass rescoring for automatic speech recognition systems. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6407–6411, 2022.

[14] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. *31st International Conference on Machine Learning, ICML 2014*, 5:1764–1772, 01 2014.

[15] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278, 2013.

[16] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[17] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[18] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, and Tara Sainath. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29:82–97, November 2012.

[19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[20] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021.

[21] Jacob Kahn, Ann Lee, and Awni Hannun. Self-training for end-to-end speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088, 2020.

[22] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE, 2020.

[23] Jian Kang, Wei-Qiang Zhang, and Jia Liu. Gated recurrent units based hybrid acoustic models for robust speech recognition. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2016.

[24] Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. Learning robust and multilingual speech representations. *arXiv preprint arXiv:2001.11128*, 2020.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[26] B. Kingsbury, Tara Sainath, and H. Soltau. Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 1:10–13, 01 2012.

[27] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

[28] Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. Rethinking evaluation in asr: Are our models robust enough? *arXiv preprint arXiv:2010.11745*, 2020.

[29] Zhiyun Lu, Yongqiang Wang, Yu Zhang, Wei Han, Zhehuai Chen, and Parisa Haghani. Unsupervised data selection via discrete speech representation for asr. In *INTERSPEECH 2022*, pages 3393–3397, 09 2022.

[30] Ji Ming and Francis Jack Smith. Improved phone recognition using bayesian triphone models. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 1:409–412 vol.1, 1998.

[31] Abdel-rahman Mohamed, George Dahl, Geoffrey Hinton, et al. Deep belief networks for phone recognition. In *NIPS workshop on deep learning for speech recognition and related applications*, volume 1, page 39, 2009.

[32] S. Nivetha. A survey on speech feature extraction and classification techniques. In *2020 International Conference on Inventive Computation Technologies (ICICT)*, pages 48–53, 2020.

[33] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[34] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[35] Chanho Park, Rehan Ahmad, and Thomas Hain. Unsupervised data selection for speech recognition with contrastive loss ratios. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8587–8591. IEEE, 2022.

[36] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[37] Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le. Improved noisy student training for automatic speech recognition. *arXiv preprint arXiv:2005.09629*, 2020.

[38] Vishal Passricha and Rajesh Kumar Aggarwal. A hybrid of deep cnn and bidirectional lstm for automatic speech recognition. *Journal of Intelligent Systems*, 29(1):1261–1274, 2020.

[39] Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[40] Mikolaj Pudo, Natalia Szczepanek, Bozena Lukasiak, and Artur Janicki. Semi-supervised learning with limited data for automatic speech recognition. *2022 IEEE 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, pages 136–141, 2022.

[41] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. Technical report, Tech. Rep., Technical report, OpenAI, 2022.

[42] Sourabh Ravindran, C. Demirogulu, and D.V. Anderson. Speech recognition using filter-bank features. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1900–1903 Vol.2, 2003.

[43] Steve Renals, Nelson Morgan, Hervé Bourlard, Michael Cohen, and Horacio Franco. Connectionist probability estimators in hmm speech recognition. *IEEE transactions on speech and audio processing*, 2(1):161–174, 1994.

[44] Tara N. Sainath, Abdel rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. Deep convolutional neural networks for lvcsr. *2013*

IEEE International Conference on Acoustics, Speech and Signal Processing, pages 8614–8618, 2013.

[45] Frank Seide, Gang Li, Xie Chen, and Dong Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *ASRU 2011*. IEEE, December 2011.

[46] Frank Seide, Gang Li, and Dong Yu. Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*. International Speech Communication Association, August 2011.

[47] Atma Prakash Singh, Ravindra Nath, and Santosh Kumar. A survey: Speech recognition approaches and techniques. In *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–4, 2018.

[48] Prakhar Swarup, Debmalya Chakrabarty, Ashtosh Sapru, Hitesh Tulsiani, Harish Arsikere, and Sri Garimella. Knowledge distillation and data selection for semi-supervised learning in ctc acoustic models. *arXiv preprint arXiv:2008.03923*, 2020.

[49] Oriol Vinyals, Suman V. Ravuri, and Daniel Povey. Revisiting recurrent neural networks for robust asr. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4085–4088, 2012.

[50] Yongqiang Wang, Abdelrahman Mohamed, Duc Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, Christian Fuegen, Geoffrey Zweig, and Michael L. Seltzer. Transformer-based acoustic modeling for hybrid speech recognition. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6874–6878, 2020.

[51] Felix Weninger, Franco Mana, Roberto Gemello, Jes'us Andr'es-Ferrer, and Puming Zhan. Semi-supervised learning with data augmentation for end-to-end asr. In *INTERSPEECH*, 2020.

[52] Shannon Wotherspoon, William Hartmann, Matthew Snover, and Owen Kimball. Improved data selection for domain adaptation in asr. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7018–7022, 2021.

[53] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020.

[54] Qiantong Xu, Tatiana Likhomanenko, Jacob Kahn, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. Iterative pseudo-labeling for speech recognition. *arXiv preprint arXiv:2005.09267*, 2020.

[55] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. A comprehensive study of deep bidirectional lstm rnns for

acoustic modeling in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2466, 2017.

[56] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

Vita

Reem Gody was born in Cairo, Egypt on 3 September 1996, the daughter of Dr. Amr M. Gody and Rania M. Salem. She received the Bachelor of Science degree in Engineering from Cairo University in 2019 and started working as a data scientist in Microsoft. She applied to the University of Texas at Austin for enrollment in their online computer science masters program. She was accepted and started graduate studies in January, 2021.

Address: reemgody@utexas.edu

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.